AD_____

Award Number:

W81XWH-12-1-0233

TITLE:
Attractor Signaling Models for Discovery of Combinatorial Therapies

PRINCIPAL INVESTIGATOR:
Carlo Piermarocchi

CONTRACTING ORGANIZATION:
Michigan State University, East Lansing, MI 48824

REPORT DATE:
September 2013

TYPE OF REPORT:
Revised Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT:

X Approved for public release; distribution unlimited

REPORT DOCUMENTATION PAGE

Form approved
OMB No. 0704-0188

1. REPORT DATE	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
September 2013	Revised Annual	15 Aug 2012 - 14 Aug 2013

4. TITLE AND SUBTITLE

Attractor Signaling Models for Discovery of Combinatorial Therapies

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-12-1-0233

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Carlo Piermarocchi

email:carlo@pa.msu.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Michigan State University
East Lansing, MI 48824

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The objective of this project consists in demonstrating that attractor models increase the chances of discovering combinations of many drugs with strong synergistic effects in lung cancer. The approach will be tested using a high throughput screening facility to compare the therapeutic effectiveness of random combinations and combinations predicted to be effective by the model.

15. SUBJECT TERMS: none listed

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON: USAMRMC**
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** UU		49	**19b. TELEPHONE NUMBER** *(include ar517 884 5631ea code)*

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

Table of Contents

Report Year 1

Grant: W81XWH-12-1-0233
PI: Carlo Piermarocchi Michigan State University
Title: Attractor signaling models for discovery of combinatorial therapies
Total period of performance: Aug 15 2012- Aug 14 2014

Introduction

Surgery and radiation therapies are difficult to use in the treatment of lung cancer because the diagnosis often occurs when patients already have metastasis. Drug-based therapies are therefore the best option, but intrinsic and acquired drug resistance still makes the 5-year survival rate for this disease less than 15%. Over the years, many specific mechanisms associated with drug resistance in lung cancer have been pinpointed, but we are still far from understanding how to overcome it. Combination drug therapy is commonly used to enhance efficacy and overcome drug resistance in cancer, but at present the choice of drugs and doses is based on empirical clinical experience alone. In this project we have used an interdisciplinary approach based on the mathematics of complex networks to identify drug combinations that could be effective in the therapy of lung cancer.

This reports describes the methods used and presents some preliminary computational and experimental data that we have obtained during the first year of operations. The project has been extended to August 14 2014 and additional details and data will be included in the final report.

Body: D*etailed description of the methods as outlined in the Statement of Work (SOW).*

1. TASK1 of SOW: Collection of data for attractor models.

A lung cell interactome was constructed by combining TRANSFAC and PhosphoPOINT data (**Subtask 1 of Task 1**). The lung network interactome we built has ~9,000 nodes and ~45,000 edges. Gene expression data was obtained from the Gene Expression Omnibus (GEO) database for A549 adenocarcinoma, H358 non-small lung cancer, and IMR90 fetal lung fibroblast normal cell lines. The model requires Boolean gene expression states. We have defined a cutoff for the normalized expression values, and all genes with expression below the cutoff are "off" and all above are "on". Because the signaling is based on a model with ± 1 states, *on* states are identified by the variable $\xi_i^a = +1$ and *off* states by $\xi_i^a = -1$, where "*a*" is either normal (*n*) or cancer (*c*).

This procedure provided the configurations corresponding to dynamical *attractor* states in our method (**Subtask 2 of Task 1**). Figure 1 shows representative gene expression data and an example of how the cut-off method was implemented.

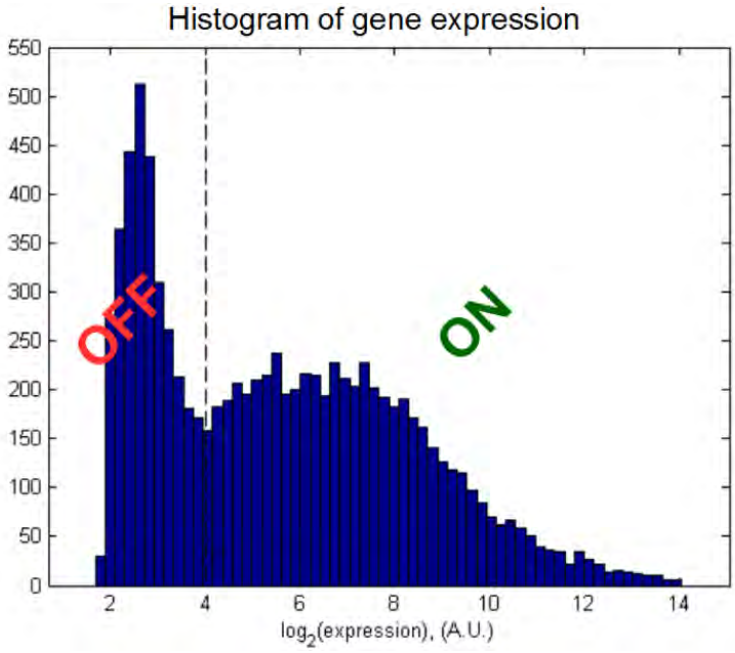

Figure 1. Representative gene expression data used in our method. Expression levels take continuous values, but must be made Boolean for our model. The expression level cutoff for normal lung cells (IMR90, pictured), for example, use a cutoff (dotted line) of approximately 4. This was chosen because the number of on states is of the same order as the number of off states, but more importantly the number of on and off states is not very sensitive to small changes in the cutoff. The same cutoff is used for both normal and cancer cells. The continuous distribution of expression levels is roughly the same for normal and cancer cells.

We have defined drug inhibitor-kinase links for a library containing about 300 kinase inhibitors using experimental surveys of kinase inhibitor targets. (**Subtask 3 of Task 1**)

2. TASK 2 of SOW: Development of attractor model based on neural network Hopfield model

After making the attractor states Boolean, we encoded the states $\vec{\xi}^{n(c)} = (\xi_1^{n(c)}, \xi_2^{n(c)}, \ldots, \xi_N^{n(c)})$

in a signaling model defined by the coupling matrix

$$J_{ij} = A_{ij}(\xi_i^n \xi_j^n + \xi_i^c \xi_j^c)\,, \qquad \text{Eq. (1)}$$

where A_{ij} is the adjacency matrix of the lung cancer network interactome obtained in Task 1, and N is the total number of nodes. The model calculates the total signal arriving at node i at time t as

$$h_i(t) = \sum_{j=1}^{N} J_{ij}\sigma_j(t),$$

where the $\sigma_i(t)$ is the state of the node i at time t. The discrete-time update scheme for the dynamical evolution of the state of the node i , $\sigma_i(t)$, is given by

$$\sigma_i(t + \Delta t) = +1 \; if \; h_i(t) > 0,$$

$$\sigma_i(t + \Delta t) = -1 \; if \; h_i(t) < 0,$$

and chosen randomly from ± 1 if the field is zero.

Note that we are left with two kinds of genes: *similarity nodes*, where $\xi_i^n = \xi_i^c$, and *differential nodes*, where $\xi_i^n = -\xi_i^c$. We have then calculated the Hamming distance between cell attractors and the dynamical state of the network (**Subtask 1 of Task 2**).

This distance has been used to identify the most sensitive single genes in the network using the following algorithm:

1. Begin with all genes set in the normal/cancer state.
2. Force gene *i=1* away from the initial state and count the number of genes that flip as a result.
3. Repeat for *i=2...N*, where *N* is the number of genes in the system.

This algorithm is effective in identifying *bottleneck* genes. *Bottlenecks* are genes which, when targeted by inhibitors, drive the cell far away from its initial state. We always try to target bottlenecks with $\xi_i^c = +1$ and $\xi_i^n = -1$ so that cancer cells are driven away from their initial state, while the normal cells are left unaltered.

We used both a one-attractor state (*p=1*) and a two-attractor state (*p=2*) signaling model. In the one attractor (*p=1*) model the J_{ij} only contains one term in Eq. (1). Both models behave like a simple Ising magnet, except that the interactions are not symmetric: the expression of gene *i* may affect the expression of gene *j*, but *j* does not necessarily affect *i*. This asymmetry makes both the *p=1* and *p=2* systems more vulnerable to external control. The *p=2* system has one property that the *p=1* doesn't, however: all edges between similarity and differential genes are effectively removed, while all edges connecting similarity genes to each other or differential genes to each other remain. The network fully separates into two independent networks, the *similarity network* and the *differential network*. When looking for nodes to target in the *p=2* case, then, all similarity nodes can be safely ignored and the problem space is significantly reduced. Aside from the edge deletion, however, *p=1* and *p=2* behave very similarly. An example of genes identified by this method and their impact *I* in terms of flipped genes in the iteractome, is shown in Table 1.

Part of the software was implemented on the high performance computer cluster facility at MSU (**Subtask 2 of Task 2**). The algorithm however was sufficiently fast that parallelization of the code was not necessary.

	I/A			
	$p=1$		$p=2$	
	Gene	I	Gene	I
UNC	HNF1A	29	OR5I1	35
	TMEM37	22	TMEM37	25
	OR5I1	20	HNF1A	23
	MAP3K14	19	POSTN	21
	MAP3K3	18	RORA	18
CON	MAP3K14	19	SRC	15
	SRC	14	BMPR1B	7

Table 1. Representative genes to be targeted for a selective killing of A549 cell line versus a control IMR90 cell line. The impact I of each gene for the p=1 and p=2 models were calculated and ranked. The constrained case (CON in the table) refers to target that are kinases and are expressed in the cancer case. The calculation is based on the selective response of I = IMR-90 (normal), A = A549 (cancer).

3. TASK 3 of SOW: First set of experiments at the high-throughput screening facility

We have carried out a first high-throughput screening of single drug and drug pair experiments (**Subtask 1 of Task 3**). The original SOW only included single drug response, but we realized that a screening with pairs would give better selectivity. 244 kinase inhibitors (KIs) of the EMD drug library were screened at 1000nM individually and the treatment lasted for 72 hours. To quantify a selective response of a cancer cell line with respect to a control normal cell line, we define the selectivity S of a single drug or drug combination as

$$S = \frac{v_N}{v_C}$$

where v_N indicates the viability of normal cells (IMR90) after treatment, and v_C the viability of cancer cells (A549) after treatment. From the screening of the 244 KIs, the top hit was PDK1/Akt1/Flt3 Dual Pathway Inhibitor (CAS # 331253-86-2) as ranked by selectivity. For the secondary screen (pair combination of drugs), we used the PDK1/Akt1/Flt3 Dual Pathway Inhibitor as the starting point and combined this compound with the other KIs as a drug pair combination. The dose of PDK1/Akt1/Flt3 Dual Pathway Inhibitor was studied to ensure proper dosing range and minimize toxicity. We used 125nM, which maintains the normal cell line IMR-90's viability >90%. For the other 243 KIs we used the standard dose of 1000nM. Several pairs in the secondary screen showed very high selectivity. The top hit from the secondary screen of the library was Alsterpaullone 2-cyanoethyl (CAS # 852529-97-0) with a selectivity of S= 6.14 for the pair (see Figure 2).

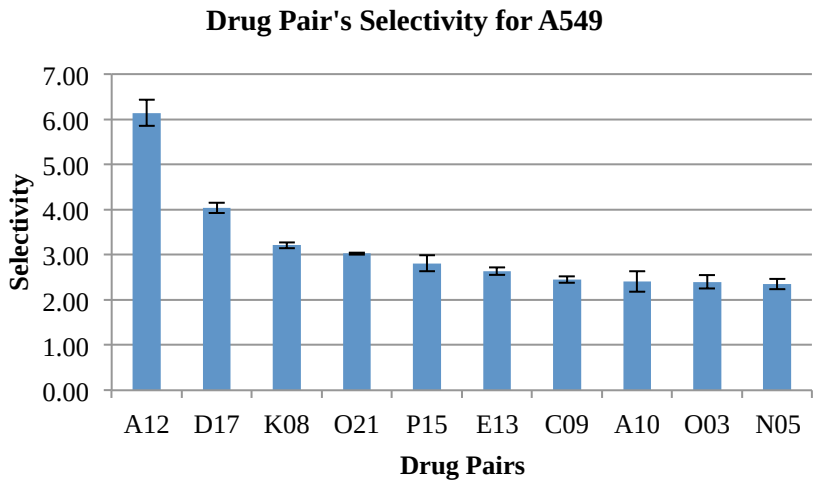

Figure 2: **Representative data from TASK 3 in SOW.** Experimental results of the top ten most selective drugs (1000nM) when paired with PDK1/Akt1/Flt3 Dual Pathway Inhibitor at 125nM. Selectivity is the IMR-90 to A549 viability ratio. The 3 digit codes identify the compounds: A12: Alsterpaullone, 2-Cyanoethyl (CAS 852529-97-0); D17: Cdk2/9 Inhibitor (CAS 507487-89-0); K08: K-252a, Nocardiopsis sp. (CAS 97161-97-2); O21: Staurosporine, Streptomyces sp. (CAS 62996-74-1); P15: WHI-P180, Hydrochloride (CAS 211555-08-7); E13: Go 6976 (CAS 136194-77-9); C09: Compound 56 (CAS 171745-13-4); A10: Alsterpaullone (CAS 237430-03-4); O03: AG 1478, Selective inhibitor of epidermal growth factor receptor (EGFR) protein (CAS 175178-82-2); N05: Reversine (CAS 656820-32-5).

We have also carried out measurements on ***random combinations*** of drugs (**Subtask 2 of Task 3**) including compounds from the EMD library and other drugs. A representative data set of random combinations is given in Table 2.

K04	A12	A15	E03	I11	628	AAG	263	MK2	662	535	Type	IMR90/A549 Selectivity
0	2	3	0	4	0	0	1	0	0	0	R	3.46
1	2	3	0	3	0	0	2	0	0	1	R	3.81
1	3	0	4	1	1	1	0	0	1	0	R	1.71
2	0	4	4	1	1	0	1	2	0	0	R	3.40
4	3	3	1	0	3	0	0	0	1	0	R	4.90
3	3	2	4	0	1	0	0	0	2	1	R	1.17
3	2	2	1	1	3	0	1	0	0	0	R	3.12
4	3	2	4	3	1	1	0	0	1	1	R	2.11
0	4	3	1	3	1	1	1	1	0	0	R	1.82
1	3	4	0	2	1	0	1	0	2	0	R	7.40

Table 2. Representative data with measurements of selectivity on A549 cells versus IMR90. Drugs were combined at different doses ranked from 0 to 4. The drug combinations obtained in this table were obtained randomly.

4. TASK 4 of SOW: Running and analysis of simulations

We have run simulations to predict the therapeutic effectiveness of combinations of kinase inhibitors according to the attractor model (**Subtask 1 of Task 4**). We have tested our model against the experimental results discussed in the previous section. We used the available kinase inhibition profiling for a drug library to determine which kinase are shut off by each drug. We applied the drugs to both the normal and cancer cells for both *p=1* and *p=2*, and compared the resulting viabilities from the experiment, v_{exp}, to the model,

$$v_{model} \sim e^{-m}$$

where "*m*" is the magnetization of the system along the attractor state (see Figure 3).

Note that the results for *p=1* and *p=2* are roughly the same, and only the *p=1* result is shown. The black circles indicate the viability of the normal cells for a given drug combination, which is the drug A15 (a PDK1/AKT1/FLT3 Inhibitor) and the drug code next to the black circles, and the connected red x's are the cancer viabilities for the same drug combination. This shows only some of the 140 drugs tested. **The most remarkable result is that without any kind of fitting, ~95% of the blue lines (including those not pictured) have a positive slope, meaning that if the experiment showed that the normal/cancer cells fared better than the cancer/normal cells, our model showed that as well.** Currently we cannot reproduce the rank of the effectiveness of the drug combinations, but we can quite accurately predict whether a combination will have a selectivity greater than or less than 1.

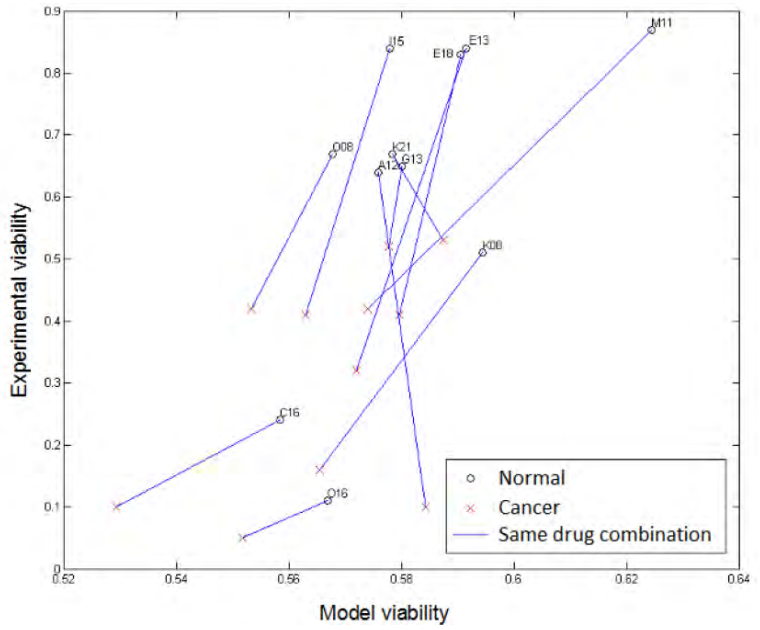

Figure 3. Computational versus experimental viability for IMR90 and A549. All drug codes shown are combined with A15 (a PDK1/AKT1/FLT3 Inhibitor). The experimental results are compared with the *p=1* model predictions (*p=2* is similar). A positive slope means that there is positive correlation between the experimental and model results: the experiment showed that normal cells treated with (A15+O16), for example, fared better than cancer cells treated with the same drugs, which our model predicts as well. Note that while only 11 drug combinations are shown, 140 were tested, a promising 95% of which had a positive slope.

We have examined combinations that are more effective using a learning machine method known as elastic net regression (**Subtask 2 of Task 4**). The method uses the in vitro lung cancer A549 cell line response of single drugs and drug pair combinations as a training set to build a regression model. Besides predicting the effectiveness of untested drugs, the method identifies sets of kinases that are statistically associated to drug sensitivity in lung cancer. More specifically, we built a regression model that predicts the response of a cell line to a drug or drug combination i. The response we predict is the normal and cancer cell viability, from which the selectivity can be derived. For this purpose, we define a regression problem in which we use the residual activity of the kinase k under the effect of drug i, which we indicate as $A_{k,i}$, as predictors of the viability. The response can be written as

$$v_i = \beta_0 + \beta_1 A_{1,i} + \cdots + \beta_p A_{p,i} \,. \qquad (2)$$

A fitting procedure based on a training set of measurements produces the coefficients $(\beta_0, \beta_1, \dots, \beta_p)$. Equation (2) can then be used to predict the viability of a new drug that has not been tested, but of which the profiling information is available. The coefficients β_k provide a measure of the sensitivity of a given cell line due to alterations in the activity of kinase k.

Subtask 3 and Subtask 4 of Task 4 are in progress and will be described in detail in the final report.

5. TASK 5 of SOW: Second set of experiments and test of hypothesis (in progress)

We are currently performing measurement of drug response of cells under combinations involving up to 10 drugs (**Subtask 1).** We have included drugs that were identified using the KIEN method above and we used a dose optimization method. Cell survival was assessed by luciferase-based assay, ATPlite™ (PerkinElmer, CA, USA), which determines viable cell numbers by measuring the presence of ATP in all metabolically active cells. For the measurement of cell viability, A549 and IMR-90 cells were plated in 384-well plates. Subsequently, the cells were treated with the drugs and 72 hours later, the ATPlite assay was performed according to the manufacturer's protocol, and luminescence was read with an Analyst HT instrument. Each combination was measured in triplicates.

Table 3 shows representative data with results of the measurements. Some of the combinations reduce the viability of cancer cells almost to zero, still significantly preserving the viability of IMR90.

K04	A12	A15	E03	I11	628	AAG	263	MK2	662	535	IMR90	Selectivity	A549
3	1	4	1	2	3	3	1	1	1	1	0.6162	297.7349	0.0021
2	2	4	2	2	3	3	1	1	1	2	0.7287	281.8844	0.0026
2	2	4	1	2	1	3	1	1	1	2	0.7257	273.3291	0.0027
2	1	4	1	2	2	3	1	1	2	2	0.6719	244.5041	0.0027
1	1	3	1	2	3	3	1	1	2	2	0.5578	225.7526	0.0025
2	2	4	2	2	3	4	1	1	1	2	0.7177	221.7178	0.0032
2	1	4	1	2	2	3	1	3	2	1	0.5110	216.8450	0.0024
2	1	4	1	2	2	3	1	1	2	1	0.5600	213.5330	0.0026
2	2	4	4	2	4	3	1	1	1	2	0.5800	210.7142	0.0028
2	1	4	3	3	3	3	1	1	1	2	0.5616	205.5397	0.0027

Table 3: Representative data with measurements of the highest selectivity on A549 cells versus IMR90. Drugs were combined at different doses ranked from 0 to 4

Conclusions

During the first year of operations, we have achieved many of the milestones defined in the statement of work. In particular we have a working code able to calculate the effect of drug combinations on the signaling of A549 adenocarcinoma, H358 non-small lung cancer, and IMR90 fibroblast normal cell lines. (Milestone 1). Two publications on the computational and theoretical results on controllability of cancer networks and identification of target genes in lung cancer have been submitted and are currently under review (Milestone 2). Experimentally, we found drugs combinations with up to 10 drugs that are very effective in killing A549 cells versus the control IMR90 cells in an in-vitro setting (Milestone 3).

Publications and Manuscripts

1) Trish Tran, Edison Ong, Andrew P Hodges, Giovanni Paternostro, Carlo Piermarocchi: *Prediction of kinase inhibitor response using activity profiling, in-vitro screening, and elastic net regression* (manuscript under review for BMC Systems Biology); attached in Appendix.
2) Anthony Szedlak, Giovanni Paternostro and Carlo Piermarocchi: *Control strategies in asymmetric Hopfield networks and application to cancer attractors* (manuscript in submitted for publication); attached in Appendix.

Prediction of kinase inhibitor response using activity profiling, *in vitro* screening, and elastic net regression.

T. Tran,[1] E. Ong,[2] A. P. Hodges,[1] G. Paternostro,[1,2] and C. Piermarocchi[2,3]

[1]*Sanford Burnham Institute for Medical Research, La Jolla, CA 92037*

[2]*Salgomed Inc., Del Mar, CA 92014*

[3]*Department of Physics and Astronomy, Michigan State University, East Lansing MI 48824*

Abstract:

Background: Many kinase inhibitors have been approved as cancer therapies. Recently, libraries of kinase inhibitors have been extensively profiled, thus providing a map of the strength of action of each compound on a large number of its targets. These profiled libraries define drug-kinase networks that can predict the effectiveness of untested drugs and elucidate the roles of specific kinases in different cellular systems. Predictions of drug effectiveness based on a comprehensive network model of cellular signalling are difficult, due to our partial knowledge of the complex biological processes downstream of the targeted kinases.

Results: We have developed the Kinase Inhibitors Elastic Net (KIEN) method, which integrates information contained in drug-kinase networks with *in vitro* screening. The method uses the *in vitro* cell response of single drugs and drug pair combinations as a training set to build linear and nonlinear regression models. Besides predicting the effectiveness of untested drugs, the KIEN method identifies sets of kinases that are statistically associated to drug sensitivity in a given cell line. We compared different versions of the method, which is based on a regression technique known as *elastic net*. Data from two-drug combinations led to predictive models, and we found that predictivity can be improved by applying logarithmic transformation to the data. The method was applied to the A549 lung cancer cell line, and we identified specific kinases known to have an important role in this type of cancer (TGFBR2, EGFR, PHKG1 and CDK4). A pathway enrichment analysis of the set of kinases identified by the method showed that axon guidance, activation of Rac, and semaphorin interactions pathways are associated to a selective response to therapeutic intervention in this cell line.

Conclusions: We have proposed an integrated experimental and computational methodology, called KIEN, that identifies the role of specific kinases in the drug response of a given cell line. The method will facilitate the design of new kinase inhibitors and the development of therapeutic interventions with combinations of many inhibitors.

Keywords: drug response predictions, kinase inhibitors, elastic net regression, high throughput screening, drug combination therapies.

1. Background

The important role of kinases in cancer biology[1] has spurred a considerable effort towards the synthesis of libraries of fully profiled kinase inhibitors, providing a map of the strength of each compound on a large number of its potential targets.[2-4] In particular, a recently published dataset has profiled several hundred kinase inhibitors using a panel of more than 300 kinases.[4] These profiled libraries define a network of interactions between drugs and their kinase targets,[5] and represent a valuable resource for the development of new therapies. In this paper, we introduce a novel computational method that incorporates profiled libraries and *in vitro* measurements to predict the response of cells to previously untested drugs. Besides making prediction about the cellular response to drugs, the method identifies critical kinase targets and pathways that are statistically associated to drug sensitivity in a given cell line.

Statistical inference and regression methods in conjunction with gene expression or mutations have been used to identify specific biomarkers associated with an increased sensitivity/resistance to drugs. For instance, the sensitivity to PARP inhibitors of Ewing's sarcoma cells with mutations in the EWS gene and to MEK inhibitors in NRAS-mutant cell lines with AHR expression have been predicted using analysis of variance and the elastic net method[6] and then experimentally validated.[7,8] In these analyses, the statistical variable associated to drugs was represented by the half maximal inhibitory concentration (IC_{50}) in different cell lines. However, besides the IC_{50}, there are many other types of information that characterize chemical compounds. These types of information can enhance the statistical analyses and improve the accuracy of predictions. For instance, a method to predict drugs sensitivity in cell lines based on the integration of genomic data with molecular physico-chemical descriptors of the drugs has been recently proposed.[9] Another useful type of information is the residual activity of kinases after interacting with a compound. Kinase profiling, patient genetic profiles, and sensitivity of primary leukemia patient samples to kinase inhibitors were recently used by Tyner *et al.*[10] to identify functionally important kinase targets and clarify kinase pathway dependence in cancer.

In this paper, the residual activity of kinases upon drug interaction is used to make predictions of the cellular response for *in vitro* experiments using an elastic net[6] regression approach. This regression method reduces the number of predictors to a minimum set, providing a clear picture of the kinases involved in the response of cell lines. A primary screen (single drug) and a secondary screen (two-drug combinations) are used as the training set for the regression. The two-drug screening exhibits a broader distribution in the response and provides a good level of predictability. In fact, the model based only on single drug response did not pass the statistical cross-validation test.

We are applying this Kinase Inhibitor Elastic Net (KIEN) method to predict cell viability of a lung cancer cell line (A549) and a normal fibroblast cell line (IMR-90) after drug treatment. We found that the regression can be improved through a logarithmic transformation on the data. Using the results of the regression, we identified a set of kinases that are strongly associated to a selective response of A549 and not IMR-90. Then, a pathway-based enrichment using Reactome[11] revealed ten significant pathways using this set of kinases, including axonal guidance and related semaphorin interactions as top hits.

This paper is organized as follows: Section 2.1 contains the experimental results of the primary and secondary *in vitro* screening corresponding to single drugs and two-drug combinations. These experimental results and residual kinase activity are analyzed with Pearson's correlation in Section 2.2. This simple correlation analysis gives a first glance of the kinases that are statistically associated to a significant change in the viability of cancer and normal cell lines. In Section 2.3, we introduce the elastic net approach and we present the results of a leave-one-out cross validation for predictions on single and pairs of drugs. We also present in this section the results obtained using the logarithmic transformation on the variables and a pathway enrichment analysis using Reactome.[11] The Discussion of the results is in Section 3, conclusions in Section 4, and Materials and Methods in Section 5.

2. Results

2.1 In vitro screen of the kinase inhibitor library

Our methodology begins with the high-throughput screening of single drug and drug pair experiments. The 244 kinase inhibitors (KIs) of the EMD drug library were screened at 1000nM individually and the treatment lasted for 72 hours. To quantify a selective response of a cancer cell line with respect to a control normal cell line, we define the selectivity S of a single drug or drug combination as

$$S = \frac{v_N}{v_C}$$

where v_N indicates the viability of normal cells (IMR90) after treatment, and v_C the viability of cancer cells (A549) after treatment. From the screening of the 244 KIs, the top hit was PDK1/Akt1/Flt3 Dual Pathway Inhibitor (CAS # 331253-86-2) as ranked by selectivity (Figure 1). For the secondary screen, we used the PDK1/Akt1/Flt3 Dual Pathway Inhibitor as the starting point and combined this compound with the other KIs as a drug pair combination. The dose of PDK1/Akt1/Flt3 Dual Pathway Inhibitor was studied to ensure proper dosing range and minimize toxicity. We used 125nM, which maintains the normal cell line IMR-90's viability

>90% (Figure 2). For the other 243 KIs we used the standard dose of 1000nM. Several pairs in the secondary screen showed very high selectivity. The top hit from the secondary screen of the library was Alsterpaullone 2-cyanoethyl (CAS # 852529-97-0) with a selectivity of S= 6.14 for the pair (Figure 3).

2.2 Analysis of correlations

In our second step, we analyzed the Pearson's correlation of the primary and secondary screening with a published dataset[4] containing target profiles for 140 kinase inhibitors. Therefore, even though we had a library of 244 KIs in the experimental screening, we were limited to utilizing 140 KIs for the analysis. For each inhibitor, the dataset provides the residual activity ($0 \leq A \leq 1$) of 291 kinases after drug treatment. This quantity is a measure of the strength of inhibition of a drug on each kinase.

For each kinase k, we calculate the Pearson's correlation, C_k, between the selectivity S_i and the activities $A_{k,i}$, with $i \in \{1, \dots, M\}$ indicating the single drug or drug pair in the set. For drug pairs, the activity is estimated as a product of the residual activities of the two drugs. The kinases are then ranked based on the p-value of their correlation with selectivity, and we calculate the False Discovery Rate (FDR) adjusted p value.[12] The list of kinases mostly correlated to the selectivity from the primary and secondary screen are listed in Table 1. We also did calculations of the correlation between the normal or cancer cell viability and the activities. The results for the top kinase-viability correlations for the primary and secondary screen are shown in the supplementary materials (Supplementary Table 1).

2.3 Elastic Net regression

Next, we build a regression model that predicts the response of a cell line to a drug or drug combination i. The response we predict is the normal and cancer cell viability, from which the selectivity can be derived. For this purpose, we define a regression problem in which we use the residual activity of the kinase k under the effect of drug i, which we indicate as $A_{k,i}$, as predictors of the viability. The response can be written as

$$v_i = \beta_0 + \beta_1 A_{1,i} + \cdots + \beta_p A_{p,i} \,. \qquad (1)$$

A fitting procedure based on a training set of measurements produces the coefficients $(\beta_0, \beta_1, \dots, \beta_p)$. Equation (1) can then be used to predict the viability of a new drug that has not been tested, but of which the profiling information is available. Note that we are integrating two different types of data: kinase profiling data is obtained through enzymatic assays that probe directly the interaction between drug and kinases, while the *in vitro* cell response data is the

result of complex signaling that involves many pathways downstream of the affected kinases. The coefficients β_k can be seen as a measure of the sensitivity of a given cell line due to alterations in the activity of kinase k.

It is well known that the least square method does not perform well in the case of linear regression with many predictors. In our case, we would like to use a database of drugs that have been profiled on about 300 kinases. However, it would be desirable to select and keep in the final model a minimal set of the kinases that provide a simple model, useful to gain biological insight. The lasso technique[13] is a powerful method to reduce the number of predictors by imposing a penalty on the regression coefficients. However, in the presence of a group of kinase predictors with strong mutual correlation, the lasso could select only one kinase predictor from the group while missing the others. To prevent this problem, our method uses the elastic net approach. This method incorporates the lasso penalty as well as a ridge penalty to keep the regression coefficients small without completely removing them.[6] The weights of the ridge and lasso penalties in the least square procedure can be optimized for best performance of the method.

We show in Figure 4 (a) and (b) the results of a leave one out cross validation (LOOCV) method for the primary (a) and secondary screen (b). For each of the 140 drugs, we apply the elastic net method using the remaining 139 drugs and then we compare the result to the measured value. This cross validation method is a particular case of the more general *k*-fold cross validation procedure in which *k* is equal to the size of the training set.[14] The cross LOOCV shows that the information contained in the primary screen is not sufficient to define a predictive model. The fact that some kinases in Table 1 show some significant correlation with the response when considered individually is in general not a sufficient condition for defining a predictive, multiple regression model. On the other hand, the secondary screen is able to reproduce the viability of many drugs, especially the ones with the stronger effect on both cell lines. Overall, the data from the secondary screen presents a much broader distribution with a tail representing a few drug combinations particularly effective. The regression works better in identifying these highly effective pairwise combinations and the relative ranking of their strengths. Data is not particularly informative for drugs and drug pair combinations that are not effective, which concentrate in the neighborhood of ~ 1 .

Data transformations can represent a powerful strategy to improve regression. We applied a logarithmic transformation, which is consistent with the hypothesis of an independent action on the different kinases on the total viability. In this case we assume that the viability can be rewritten in the form

$$v_i = e^{\beta_0} \left(A_{1,i}\right)^{\beta_1} \cdot \left(A_{2,i}\right)^{\beta_2} \cdot \ldots \cdot \left(A_{p,i}\right)^{\beta_p} . \qquad (2)$$

By applying a *log* transformation on both sides of Eq. (2) we reduce the problem to a linear regression, to which the elastic net strategy can be applied. We show in Figure 5 the results of the LOOCV for the primary and secondary screen using the logarithmic data transformation. As in the linear case, we find that the method fails the cross validation procedure if we use data from the primary screen, while the secondary screen with log transformed data gives better R^2.

In addition to a regression model that can be used to predict the efficacy of drugs that have not been tested, the β_i coefficients can be used to rank kinases in terms of their relevance in the regression. Therefore, these coefficients identify the kinases whose inhibition is associated to a decrease in the cell viability. A ranking based on the differential $\beta_i^C - \beta_i^N$, where the index *N* and *C* identify the regression model of the cancer and normal cells, gives insight on specific pathways important for a selective response of cancer cells. Table 2 shows a list of kinases ranked in terms of $\left|\beta_i^C - \beta_i^N\right|$, where the coefficients have been obtained using the logarithmic data transformation on the secondary screen.

In order to test whether selected pathways were significantly enriched for the identified kinase genes in Table 2, a pathway-based enrichment analysis was conducted using the results from the elastic net kinase analysis and Fisher exact tests. Ten pathways from Reactome were identified as significant ($p<0.05$) using this kinase list, including axon guidance, activation of Rac, and semaphorin interactions as top hits (Table 3).

3. Discussion

Drug-kinase profiling represents a controller-target network[5] that when combined with *in vitro* testing, can be used in regression models to predict drug response and to identify pathways statistically associated to drug sensitivity. Network methods in biology are often based on the analysis of large datasets from high-throughput experiments. An example is given by gene regulatory networks, which presents many challenges either when restricted to a homogeneous set of data[15,16] or when it includes different classes of data.[17-20] In our KIEN method, information from the drug-target network and experimental query of the biological system are integrated. The goal is not a reconstruction of a regulatory network, but to identify a set of kinases linked to a therapeutic response in a given cell line. In order to establish associations, the system has to be perturbed by the use of kinase inhibitor drugs. The response to these single drugs or drug combinations becomes a training set that when combined with the kinase profiling, can lead to predictions.

The elastic net method is one of the most widely used regularization techniques. Regularization techniques are used in statistical and machine learning models to achieve an optimal tradeoff between accuracy and simplicity. Simplicity makes a model less prone to

overfitting and more likely to generalize. In our analysis, we found that the elastic net regressions based on single drug responses were not successful, while drug pair data provided statistically significant predictions. A possible explanation for this finding is the following: single drugs might be less able to overcome the robustness of biological networks.[5] The phenotypic signal is therefore blunted and not easily measured. If a second drug is added, any compensatory capacity is already stretched and the effects from the inhibition of each kinase can be seen more clearly. Using data from drug pairs, we found that noise can be better filtered out and stronger statistical associations between kinases and therapeutic response are revealed. Clearly, if a different training set with higher variance in efficacy measures were used in the primary screen, it is likely that also single drug *in vitro* response would have given a significant predictive model.

We identified several kinases that are implicated in lung cancer that gives biological significance to our KIEN method. In particular, TGFBR2 appears as a top hit both in the correlation and in the elastic net methods. This finding is consistent with recent siRNA experiments on A549 cell lines,[21] which demonstrated that silencing of this receptor reduces cell proliferation, invasion, and metastasis. The Cyclin-dependent kinase 4 (CDK4) appears as a second top target in the correlation analysis, and is also highly significant in the KIEN analysis. Experiments using lentiviral-mediated shRNA to inhibit CDK4 in A549 have shown inhibited cell cycle progression, suppressed cell proliferation, colony formation, and migration,[22] and there is an ongoing clinical trial using a CDK4/6 inhibitor in lung cancer.[23] The KIEN analysis identified EGFR, which is known to be overexpressed in the majority of non-small cell lung cancers.[24] Furthermore, RNAi experiments targeting EGFR demonstrated cancer growth suppression in A549 xenograft in mice.[25] The third kinase in Table 2, PHKG1 has also been found to be upregulated in human tumor samples, including lung adenocarcinoma, and aberrations in its gene copy number is a feature of many human tumors[26].

The pathway-based enrichment provides a broader view on the role of the kinases identified by our method in Table 2. Among the top three pathways shown in Table 3 are activation of Rac and Semaphorin interactions. Rac proteins play a key role in cancer signaling and they belong to the RAS superfamily.[27] We also identified a set of semaphorins in our analysis that is represented in the top significantly enriched pathways. Semaphorins, previously known as collapsins, are a set of proteins containing a 500-amino acid sema domain among others (including PSI and immunoglobulin type domains), which can be transmembranous or secreted.[28] It is known that Sema3E cleavage promotes invasive growth and metastasis *in vivo*.[28] These genes also have selective targeting by Rac and Rho family members. This generates hypotheses of possible pathways that could be targeted therapeutically. However, these

hypotheses need to be validated by further experiments with different inhibitors for the same targets or with alternative methods, e.g. using siRNA.

4. Conclusions

We have introduced an integrated experimental and computational methodology that identifies the role of specific kinases in the drug response of a given cell line. The key element of our KIEN methodology is a multiple regression procedure that uses *in vitro* screen data as a training set. If a new library of kinase inhibitor compounds were to be synthetized and profiled, then our model would be able to immediately estimate the effect of these drugs on *in vitro* experiments on a given cell line. We have shown an application to a lung cancer cell line, but our method can be extended to different cell lines. The method will facilitate the design of new kinase inhibitors and the development of therapeutic interventions with combinations of many inhibitors.[29] The procedure could be extended to three drug combinations, if measurements for these larger combinations were available. Finally, the method could be extended to regression models that are specific of cancer cells with the same set of mutations, or it could be directly used with patient-derived primary cells to identify a personalized treatment.

5. Materials and Methods

Materials

The primary screening of a kinase inhibitor (KI) library comprised of 244 KIs was purchased from EMD Chemicals, and diluted with DMSO to 2mM concentrations for high-throughput screening purposes. The KI library was stored at -80°C. Additionally, PDK1/Akt1/Flt3 Dual Pathway Inhibitor (CAS # 331253-86-2) was ordered from EMD. Only 140 out of 244 were used in the drug-target network reconstruction because the drug profiling information was available only for these compounds. One kinase inhibitor known to affect the kinase targets indirectly was excluded.

Cell Culture

Cell lines IMR-90 (normal lung fibroblast) and A549 (lung adenocarcinoma) were cultured in RPMI 1640 (Hyclone) supplemented with 10% Canadian characterized fetal bovine serum (Hyclone), 1% 200mM L-glutamine (Omega), and 1% penicillin/streptomycin (Omega). The media for the cells were renewed every 3 days and kept at 80-90% confluency. Cells were maintained in a humidified environment at 37°C and 5% CO_2.

Kinase Inhibitor Experiments

IMR-90 (1500 cells/well) and A549 (750 cells/well) were seeded on 384-well microplates (Grenier Bio-One) and incubated for 3 hours before the addition of kinase inhibitor(s). The reason that IMR-90 was seeded at double the cell density of A549 is due to the difference in cell division. IMR-90's doubling time is 36-48 hours whereas A549's is 22 hours. We wanted to make sure that the cells have divided at least once during the 72hr drug treatment. Furthermore, both A549's and IMR-90's final confluency at 72 hrs is 90-95% and within the range of the ATPlite 1step assay. Supplementary Figures S1 and S2 show the growth curve for both cell lines. IMR-90 and A549 cell lines were tested on the same day with three replicates and the experiment was repeated three times with randomized well positions to reduce biases. ECHO 555 Liquid Handler (Labcyte) was used to dispense nanoliter volumes of each KI to 384-well plates with cells attached (wet dispense). The final volume in the plate is 40uL and cells were incubated for 72 hours with KI treatment.

ATP Measurements

ATPlite 1Step (Perkin Elmer) was used to evaluate the cell number and cytotoxicity. ATP measurements were done by dispensing 20uL of the ATPlite 1Step solution to each well to a final volume of 60uL. The plate was placed on a shaker at 1100rpm and the luminescence activity was detected by Analyst GT Plate Reader. The percent (%) of control is the quantity of ATPlite 1step measurement of the treated versus the untreated wells of each individual cell type. The ATP standard was prepared with culture media to final volume of 40uL, and 20uL of ATPlite 1step reagent was added. Supplementary Figure 3 shows the ATP standard curve. The plate was read immediately.

Computational Methods

Correlations between selectivity/viability and kinase activity were calculated using the python *scipy linregress* function, which also provide *p-values.* Ranking the p-values and directly applying the Benjamini–Hochberg procedure gave us the FDR values. The elastic net regression was carried out using the Scikit-learn package[30] which finds the coefficients β that minimize the function

$$F = \frac{1}{2\,M}||v - A\beta||_2^2 + \alpha\rho||\beta||_1 + \frac{1}{2}\alpha(1-\rho)||\beta||_2^2\,,$$

where v is the vector of the observed viabilities and A is the matrix containing the residual activity of the kinases from the profiling, and M is the total number of drugs or drug

combinations used. The parameters α and β determine the relative weights of the lasso and ridge penalties quantified using L^1 ($||\cdot||_1$) and L^2 ($||\cdot||_2$) norm, respectively. We used $\alpha = 0.15$ and $\rho = 0.01$ in the results of Figures 4 and 5 and in Table 2. We also tried other values of these parameters, which did not give a significant difference in the results.

Pathway-based enrichment

Reactome pathways were downloaded using a newer build of the 'biomaRt' library (v2.12.0) in Bioconductor /R (v2.15.0). Gene symbols from the kinase list were converted to Entrez gene identifier numbers ('entrezgene') and mapped against the gene ids in each Reactome pathway. For each pathway, the set of significant genes enriched within any given pathway was computed using a Fisher exact test. The procedure computes the significance (p-value) of observing significant kinases, as deemed significant by our method, within the selected pathway. These pathways are identified from 518 Reactome pathways. Given that our gene set consists entirely of kinases and would be generalized towards kinase-specific effects, the set of all kinases (~300) were selected for background adjustment and more sensitive enrichment of the pathways. This procedure was repeated for each pathway to generate p-values and pathway rankings. False discovery rate [FDR] values were later generated to further restrict significance.

6. Acknowledgements

We thank Dr. Anthony (Tony) Polverino for many discussions. We would like to acknowledge the NSF grant (CCF0829891) and the DOD/CDMRP Lung Cancer Research Program (grant W81XWH-12-1-0233) for support.

The data sets supporting the results of this article are included within the article. Authors' contributions: GP & CP proposed concept, EO, CP and APH performed calculations, TT performed experiment, CP and GP wrote the manuscript. Correspondence and requests for materials should be addressed to giovanni@sanfordburnham.org or carlo@pa.msu.edu.

FIGURES

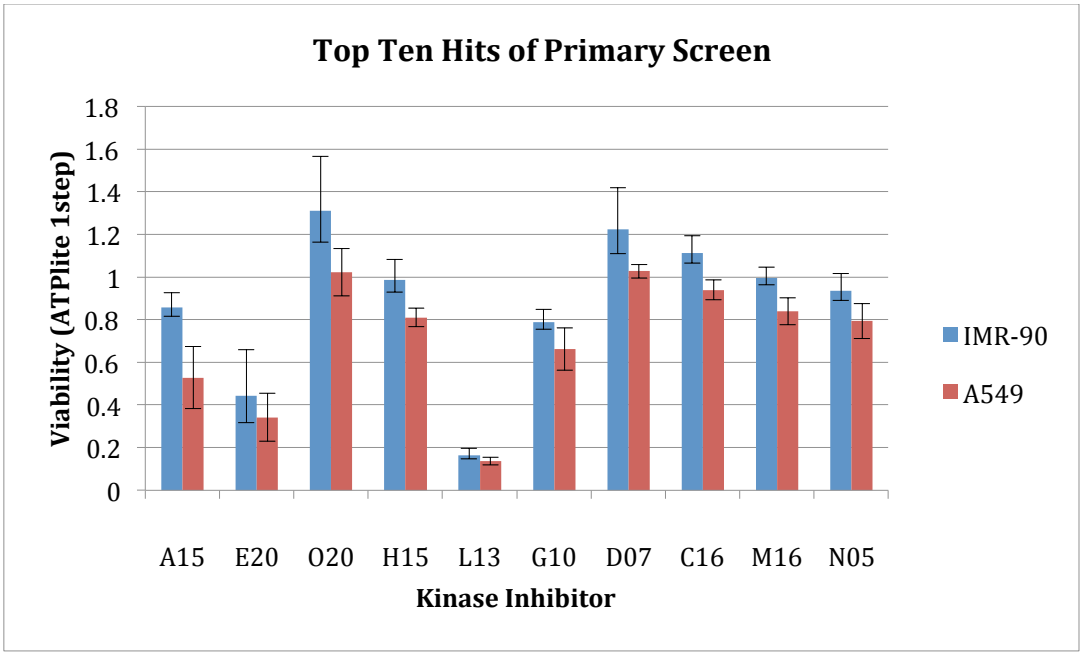

Figure 1. Primary screen results of the top ten most selective kinase inhibitors. Drugs are ranked based on the IMR-90 to A549 viability ratio. The 3 digit codes identify the compounds: A15: PDK1/Akt1/Flt3 Dual Pathway Inhibitor (CAS 331253-86-2); E20: Cdk/Crk Inhibitor (CAS 784211-09-2); O20: SU9516 (CAS 666837-93-0); H15: MEK1/2 Inhibitor II (CAS 212631-61-3); L13: PI 3-Kα Inhibitor VIII (CAS 372196-77-5); G10: Fascaplysin, Synthetic (CAS 114719-57-2); D07: Cdk2 Inhibitor II (CAS 222035-13-4); C16: Cdk1/2 Inhibitor III (CAS 443798-55-8); M16: GSK3b Inhibitor XII, TWS119 (CAS 601514-19-6); N05: Reversine (CAS 656820-32-5). The chemical structure of these compounds is given in a supplementary file.

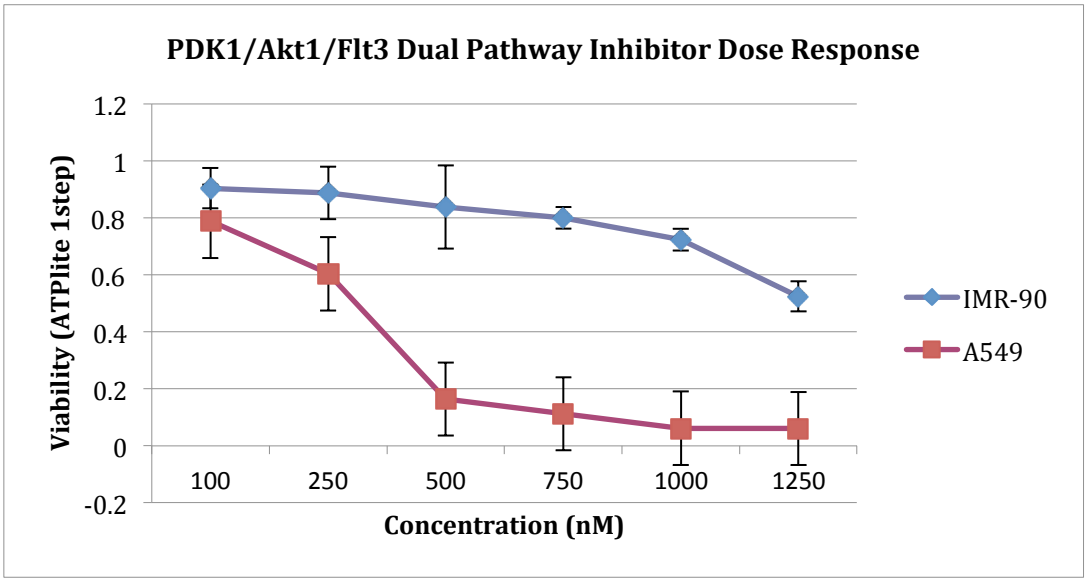

Figure 2**.** Dose response curve of PDK1/Akt1/Flt3 Dual Pathway Inhibitor. Different doses of PDK1/Akt1/Flt3 Dual Pathway Inhibitor were tested to measure the response of A549 to the drug. For the secondary screen we selected 125nM to ensure low toxicity on the normal cell line.

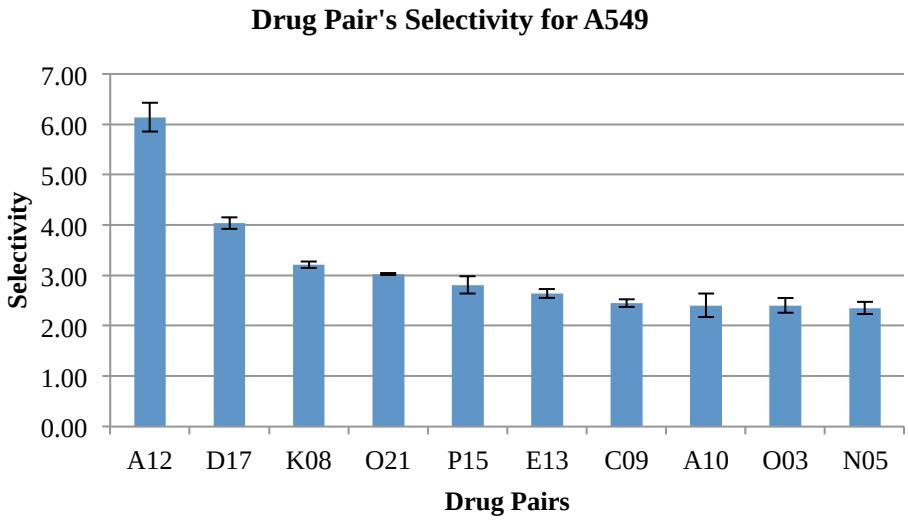

Figure 3. Secondary screen results of the top ten most selective drugs (1000nM) when paired with PDK1/Akt1/Flt3 Dual Pathway Inhibitor at 125nM. Selectivity is the IMR-90 to A549 viability ratio, as defined in Section 2.1. The 3 digit codes identify the compounds: A12: Alsterpaullone, 2-Cyanoethyl (CAS 852529-97-0); D17: Cdk2/9 Inhibitor (CAS 507487-89-0); K08: K-252a, Nocardiopsis sp. (CAS 97161-97-2); O21: Staurosporine, Streptomyces sp. (CAS 62996-74-1); P15: WHI-P180, Hydrochloride (CAS 211555-08-7); E13: Gö 6976 (CAS 136194-77-9); C09: Compound 56 (CAS 171745-13-4); A10: Alsterpaullone (CAS 237430-03-4); O03: AG 1478, Selective inhibitor of epidermal growth factor receptor (EGFR) protein (CAS 175178-82-2); N05: Reversine (CAS 656820-32-5). The chemical structure of these compounds is given in a supplementary file.

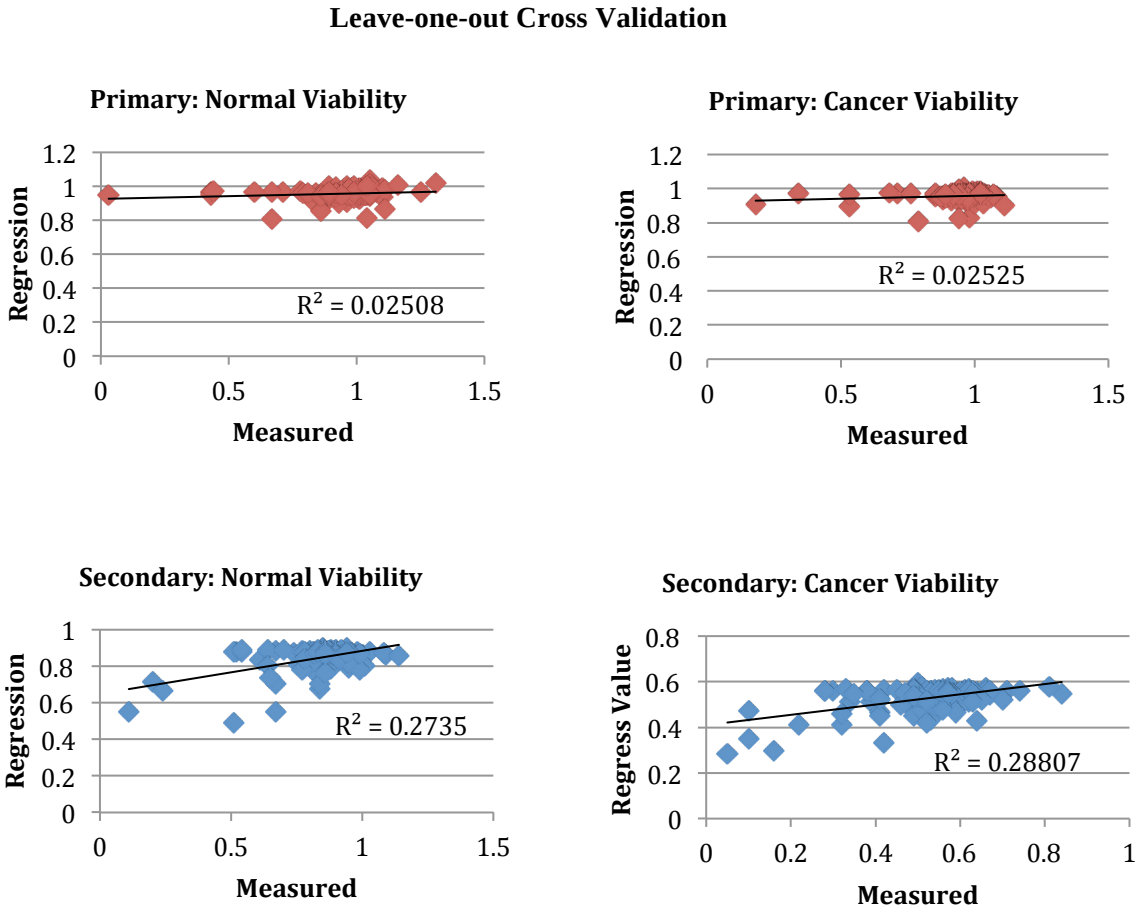

Figure 4: Leave-one-out Cross Validation of the elastic net regression model based on the primary (top) and secondary (bottom) screens for normal and cancer cell lines. Each of the 140 point in these figures corresponds to one of the 140 drug. "Regression" refers to the viability predicted by the regression model using all data from the other 139 drugs as training set, while "Measured" refers to the actual viability measured for the drug or drug combination. Note that only the secondary screen leads to predictive models with significant R^2 for the two cancer cell types.

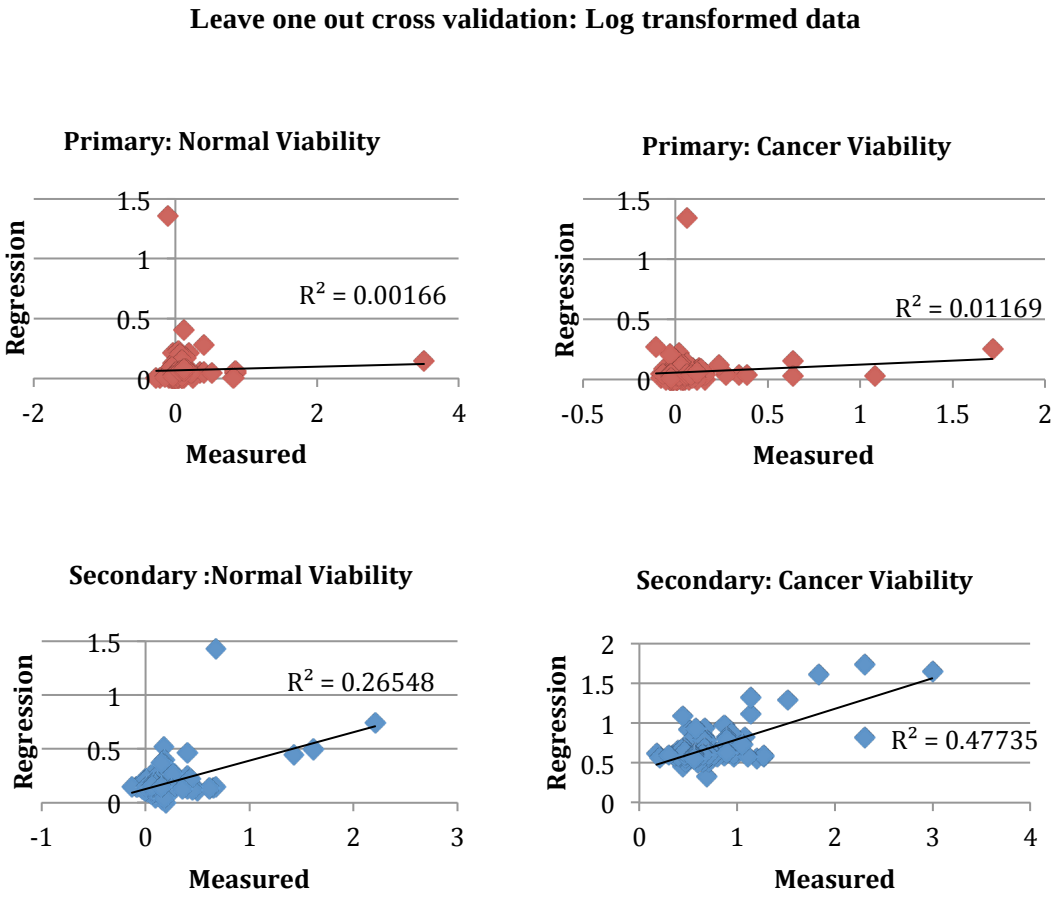

Figure 5: Leave-one-out Cross Validation of the elastic net regression model based on the primary (top) and secondary (bottom) screens for normal and cancer cell lines after logarithmic transformation on the data. Each of the 140 point in these figures corresponds to one of the 140 drugs. "Regression" refers to *–log* of the viability predicted by the regression model using all data from the other 139 drugs as training set, while "Measured" refers to *–log* of the actual viability measured for the drug or drug combination. Note that, as in Figure 4, only the secondary screen leads to predictive models with significant R^2 for both cell types. The R^2 for the Cancer cell lines is considerably better using the log transformation.

TABLES

Kinase	Selectivity Corr	FDR		Kinase	Selectivity Corr	FDR
Primary screening				Secondary Screening		
PRKCZ	0.451	2.28E-08		TGFBR2	-0.501	8.29E-08
DMPK	0.435	7.75E-08		CDK4	-0.412	6.40E-05
STK39	0.430	1.15E-07		CDC42BPB	-0.409	6.40E-05
EPHA8	0.420	2.33E-07		RIPK2	-0.399	7.73E-05
ADRBK2	0.399	1.01E-06		DSTYK	-0.369	0.000413
PRKACG	0.396	1.27E-06		ACVRL1	-0.368	0.000413
CAMK4	0.394	1.45E-06		PAK1	-0.367	0.000413
MAP2K2	0.393	1.53E-06		MAPKAPK2	-0.364	0.000413
ADRBK1	0.392	1.62E-06		PAK7	-0.359	0.000424
PNCK	0.382	3.29E-06		CDK1	-0.357	0.000429

Table 1 Correlations between selectivity and kinase activity from primary and secondary screening. A negative correlation indicates that inhibition of that particular kinases is associated to a higher selectivity. The top two hits with negative correlation, TGFBR2 and CDK4 are known to have an important role in cell proliferation, invasion and metastasis in lung adenocarcinoma[21,22].

Kinase	Cancer beta Coefficient	Normal beta Coefficient	Difference
TGFBR2	0.061	0.000	0.061
EGFR	0.060	0.000	0.060
PHKG1	0.051	0.014	0.037
RIPK2	0.032	-0.002	0.034
PRKG2	0.012	0.045	0.033
CDK4	0.021	-0.008	0.029
MAP3K10	0.038	0.014	0.024
MARK4	0.000	0.022	0.022
PAK1	0.025	0.004	0.021
MAP4K5	0.021	0.000	0.021
MARK2	0.006	0.026	0.021
MARK3	0.000	0.020	0.020
TBK1	0.012	0.031	0.020
ERBB2	0.021	0.001	0.019
NUAK1	-0.029	-0.010	0.019
ULK2	0.018	0.000	0.018
MYLK2	-0.024	-0.006	0.018
MAP4K4	0.004	-0.014	0.018
CDK5	0.002	-0.016	0.018
GSK3B	0.021	0.004	0.017
PAK2	0.019	0.002	0.017
CDC42BPB	0.023	0.006	0.017
DSTYK	0.006	-0.010	0.016
RPS6KA2	0.000	-0.016	0.016
FGFR1	-0.004	0.012	0.016
PAK7	0.015	0.000	0.015
PIM1	-0.015	0.000	0.015
CDK3	0.015	0.000	0.015
IRAK1	-0.002	-0.017	0.015

Table 2. Kinases with the highest difference in the regression coefficients for the log transformed data of the secondary screen. A larger difference is associated with a selective response of A549 upon inhibition. Note that in addition to TGFB2R and CDK4, which were identified with the correlation approach of Table 1, additional kinases known to have an important role in lung cancer such as EGFR [24,25] and PHKG1[26] are found using the elastic net approach.

Path ID	Path name	N_S	N_T	p-val
422475	Axon guidance	9	31	0.005
428540	Activation of Rac	3	5	0.008
373755	Semaphorin interactions	4	10	0.011
376176	Signaling by Robo receptor	3	7	0.024
1266738	Developmental Biology	8	39	0.026
445144	Signal transduction by L1	4	13	0.030
373760	L1CAM interactions	4	14	0.040
193639	p75NTR signals via NF-kB	2	4	0.051
209543	p75NTR recruits signaling complexes	2	4	0.051
389359	CD28 dependent Vav1 pathway	2	4	0.051

Table 3. Reactome pathways with significant representation of kinases from the regression analysis. N_s indicates the number of kinases that are found significant in the regression analysis, while N_T is the total number of kinases in the pathway. The top ten pathways with Fisher exact test $p<=0.051$ are shown. These pathways are identified from 518 Reactome pathways containing at least one of the kinases identified in Table 2. The 9 kinases in the axon-guidance pathway are EGFR, PAK1, ERBB2, CDK5, GSK3B, PAK2, RPS6KA2, FGFR1 and PAK7.

REFERENCES:

1 Cohen, P. Protein kinases - the major drug targets of the twenty-first century? *Nat Rev Drug Discov* **1**, 309-315, (2002).

2 Fabian, M. A., Biggs, W. H., Treiber, D. K., Atteridge, C. E., Azimioara, M. D., Benedetti, M. G., Carter, T. A., Ciceri, P., Edeen, P. T., Floyd, M., Ford, J. M., Galvin, M., Gerlach, J. L., Grotzfeld, R. M., Herrgard, S., Insko, D. E., Insko, M. A., Lai, A. G., Lelias, J.-M., Mehta, S. A., Milanov, Z. V., Velasco, A. M., Wodicka, L. M., Patel, H. K., Zarrinkar, P. P. & Lockhart, D. J. A small molecule-kinase interaction map for clinical kinase inhibitors. *Nat Biotech* **23**, 329-336, (2005).

3 Karaman, M. W., Herrgard, S., Treiber, D. K., Gallant, P., Atteridge, C. E., Campbell, B. T., Chan, K. W., Ciceri, P., Davis, M. I., Edeen, P. T., Faraoni, R., Floyd, M., Hunt, J. P., Lockhart, D. J., Milanov, Z. V., Morrison, M. J., Pallares, G., Patel, H. K., Pritchard, S., Wodicka, L. M. & Zarrinkar, P. P. A quantitative analysis of kinase inhibitor selectivity. *Nat Biotech* **26**, 127-132, (2008).

4 Anastassiadis, T., Deacon, S. W., Devarajan, K., Ma, H. C. & Peterson, J. R. Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. *Nat Biotechnol* **29**, 1039-U1117, (2011).

5 Feala, J. D., Cortes, J., Duxbury, P. M., McCulloch, A. D., Piermarocchi, C. & Paternostro, G. Statistical Properties and Robustness of Biological Controller-Target Networks. *PLoS ONE* **7**, e29374, (2012).

6 Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 301-320, (2005).

7 Garnett, M. J., Edelman, E. J., Heidorn, S. J., Greenman, C. D., Dastur, A., Lau, K. W., Greninger, P., Thompson, I. R., Luo, X. & Soares, J. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570-575, (2012).

8 Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehar, J., Kryukov, G. V. & Sonkin, D. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-607, (2012).

9 Menden, M. P., Iorio, F., Garnett, M., McDermott, U., Benes, C. H., Ballester, P. J. & Saez-Rodriguez, J. Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. *PLoS ONE* **8**, (2013).

10 Tyner, J. W., Yang, W. F., Bankhead, A., Fan, G., Fletcher, L. B., Bryant, J., Glover, J. M., Chang, B. H., Spurgeon, S. E. & Fleming, W. H. Kinase pathway dependence in primary human leukemias determined by rapid inhibitor screening. *Cancer research* **73**, 285-296, (2013).

11 Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H. & Jassal, B. Reactome knowledgebase of human biological pathways and processes. *Nucleic acids research* **37**, D619-D622, (2009).

12 Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289-300, (1995).
13 Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288, (1996).
14 Kohavi, R. in *International joint Conference on artificial intelligence.* 1137-1145 (Lawrence Erlbaum Associates Ltd, 1995).
15 De Smet, R. & Marchal, K. Advantages and limitations of current network inference methods. *Nature Reviews Microbiology* **8**, 717-729, (2010).
16 Marbach, D., Prill, R. J., Schaffter, T., Mattiussi, C., Floreano, D. & Stolovitzky, G. Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the National Academy of Sciences* **107**, 6286, (2010).
17 Marbach, D., Costello, J. C., Kuffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., Allison, K. R., Kellis, M., Collins, J. J., Stolovitzky, G. & Consortium, D. Wisdom of crowds for robust gene network inference. *Nature Methods* **9**, 796-+, (2012).
18 Bar-Joseph, Z., Gerber, G. K., Lee, T. I., Rinaldi, N. J., Yoo, J. Y., Robert, F., Gordon, D. B., Fraenkel, E., Jaakkola, T. S. & Young, R. A. Computational discovery of gene modules and regulatory networks. *Nat Biotechnol* **21**, 1337-1342, (2003).
19 Lemmens, K., De Bie, T., Dhollander, T., De Keersmaecker, S. C., Thijs, I. M., Schoofs, G., De Weerdt, A., De Moor, B., Vanderleyden, J. & Collado-Vides, J. DISTILLER: a data integration framework to reveal condition dependency of complex regulons in Escherichia coli. *Genome Biol* **10**, R27, (2009).
20 Reiss, D., Baliga, N. & Bonneau, R. Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC bioinformatics* **7**, 280, (2006).
21 Xu, C. C., Wu, L. M., Sun, W., Zhang, N., Chen, W. S. & Fu, X. N. Effects of TGF-beta signaling blockade on human A549 lung adenocarcinoma cell lines. *Mol Med Rep* **4**, 1007-1015, (2011).
22 Wu, A. B., Wu, B., Guo, J. S., Luo, W. R., Wu, D., Yang, H. L., Zhen, Y., Yu, X. L., Wang, H., Zhou, Y., Liu, Z., Fang, W. Y. & Yang, Z. X. Elevated expression of CDK4 in lung cancer. *J Transl Med* **9**, (2011).
23 http://clinicaltrials.gov. <http://clinicaltrials.gov/show/NCT01291017> (2014).
24 Brabender, J., Danenberg, K. D., Metzger, R., Schneider, P. M., Park, J. M., Salonga, D., Holscher, A. H. & Danenberg, P. V. Epidermal growth factor receptor and HER2-neu mRNA expression in non-small cell lung cancer is correlated with survival. *Clin Cancer Res* **7**, 1850-1855, (2001).
25 Li, C., Zhang, X., Cheng, L., Dai, L., Xu, F., Zhang, J., Tian, H., Chen, X., Shi, G., Li, Y., Du, T., Zhang, S., Wei, Y. & Deng, H. RNA interference targeting human FAK and EGFR suppresses human non-small-cell lung cancer xenograft growth in nude mice. *Cancer Gene Therapy* **20**, 101+, (2013).
26 Camus, S., Quevedo, C., Menendez, S., Paramonov, I., Stouten, P. F. W., Janssen, R. A. J., Rueb, S., He, S., Snaar-Jagalska, B. E., Laricchia-Robbio, L. & Belmonte, J. C. I. Identification of phosphorylase kinase as a novel therapeutic target through high-throughput screening for anti-angiogenesis compounds in zebrafish. *Oncogene* **31**, 4333-4342, (2012).

27 Kawazu, M., Ueno, T., Kontani, K., Ogita, Y., Ando, M., Fukumura, K., Yamato, A., Soda, M., Takeuchi, K., Miki, Y., Yamaguchi, H., Yasuda, T., Naoe, T., Yamashita, Y., Katada, T., Choi, Y. L. & Mano, H. Transforming mutations of RAC guanosine triphosphatases in human cancers. *P Natl Acad Sci USA* **110**, 3029-3034, (2013).

28 Potiron, V. A., Roche, J. & Drabkin, H. A. Semaphorins and their receptors in lung cancer. *Cancer Lett* **273**, 1-14, (2009).

29 Feala, J. D., Cortes, J., Duxbury, P. M., Piermarocchi, C., McCulloch, A. D. & Paternostro, G. Systems approaches and algorithms for discovery of combinatorial therapies. *Wires Syst Biol Med* **2**, 181-193, (2010).

30 Pedregosa, F., Varoquaux, G. l., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R. & Dubourg, V. Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research* **12**, 2825-2830, (2011).

Control of asymmetric Hopfield networks and application to cancer attractors

Anthony Szedlak[1], Giovanni Paternostro[2], Carlo Piermarocchi[1]
[1] *Department of Physics and Astronomy, Michigan State University, East Lansing MI 48824*
[2] *Sanford-Burnham Medical Research Institute, 10901 North Torrey Pines Road, La Jolla, CA 92037*
(Dated: January 31, 2014)

The asymmetric Hopfield model is used to simulate signaling dynamics in gene/transcription factor networks. The model allows for a direct mapping of a gene expression pattern into attractor states. We analyze different control strategies aiming at disrupting attractor patterns using selective local fields representing therapeutic interventions. The control strategies are based on the identification of signaling *bottlenecks*, which are single nodes or strongly connected clusters of nodes that have a large impact on the signaling. We provide a theorem with bounds on the minimum number of nodes that guarantee controllability of bottlenecks consisting of strongly connected components. The control strategies are applied to the identification of sets of proteins that, when inhibited, selectively disrupt the signaling of cancer cells while preserving the signaling of normal cells. We use an experimentally validated non-specific network and a specific B cell interactome reconstructed from gene expression data to model cancer signaling in lung and B cells, respectively. This model could help in the rational design of novel robust therapeutic interventions based on our increasing knowledge of complex gene signaling networks.

PACS numbers: 87.16.A-,87.16.Xa

I. INTRODUCTION

The vision behind systems biology is that complex interactions and emergent properties determine the behavior of biological systems. Many theoretical tools developed in the framework of spin glass models are well suited to describe emergent properties, and their application to large biological networks represents an approach that goes beyond pinpointing the behavior of a few genes or metabolites in a pathway. The Hopfield model [1] is a spin glass model that was introduced to describe neural networks, and that is solvable using mean field theory [2]. The asymmetric case, in which the interaction between the spins can be seen as directed, can also be exacty solved in some limits [3]. The model belongs to the class of attractor neural networks, in which the spins evolve towards stored attractor patterns, and it has been used to model biological processes of high current interest, such as the reprogramming of pluripotent stem cells [4]. Moreover, it has been suggested that a biological system in a chronic or therapy-resistant disease state can be seen as a network that has become trapped in a pathological Hopfield attractor [5]. A similar class of models is represented by Random Boolean Networks [6], which were proposed by Kauffman to describe gene regulation and expression states in cells [7]. Differences and similarities between the Kauffman-type and Hopfield-type random networks have been studied for many years [8–11].

In this paper, we consider an asymmetric Hopfield model built from realistic (even if incomplete [12, 13]) cellular networks, and we map the spin attractor states to gene expression data from normal and cancer cells. We will focus on the question of the control of the dynamical properties of the network using external local fields representing therapeutic interventions. To a major extent, the final determinant of cellular phenotype is the expression and activity pattern of all proteins within the cell, which is related to levels of mRNA transcripts. Microarrays measure genome-wide levels of mRNA expression that therefore can be considered a rough snapshot of the state of the cell. This state is relatively stable, reproducible, unique to cell types, and can differentiate cancer cells from normal cells, as well as differentiate between different types of cancer [14, 15]. In fact, there is evidence that attractors exist in gene expression states, and that these attractors can be reached by different trajectories rather than only by a single transcriptional program [16]. While the dynamical attractors paradigm has been originally proposed in the context of cellular developement, the similarity between cellular *ontogenesis*, i.e. the developement of different cell types, and *oncogenesis*, i.e. the process under which normal cells are transformed into cancer cells, has been recently emphasized [17]. The main hypothesis of this paper is that cancer robustness is rooted in the dynamical robustness of signaling in an underlying cellular network. If the cancerous state of rapid, uncontrolled growth is an attractor state of the system [18], a goal of modeling therapeutic control could be to design complex therapeutic interventions based on drug combinations [19] that push the cell out of the cancer attractor basin. [20]

Many authors have discussed the control of biological signaling networks using complex external perturbations. Calzolari and coworkers considered the effect of complex external signals on apoptosis signaling [21]. Agoston and coworkers [22] suggested that perturbing a complex biological network with partial inhibition of many targets could be more effective than the complete inhibition of a single target, and explicitly discussed the implications for multi-drug therapies [23]. In the traditional approach to control theory [24], the control of a dynamical system consists in finding the specific input temporal se-

quence required to drive the system to a desired output. This approach has been discussed in the context of Kauffmann Boolean networks [25] and their attractor states [26]. Several studies have focused on the intrinsic global properties of control and hierarchical organization in biological networks [27, 28]. A recent study has focused on the minimum number of nodes that needs to be addressed to achieve the complete control of a network [29]. This study used a linear control framework, a matching algorithm [30] to find the minimum number of controllers, and a replica method to provide an analytic formulation consistent with the numerical study. Finally, Cornelius *et al.* [31] discussed how nonlinearity in network signaling allows reprogrammig a system to a desired attractor state even in the presence of contraints in the nodes that can be accessed by external control. This novel concept was explicitly applied to a T-cell survival signaling network to identify potential drug targets in T-LGL leukemia. The approach in the present paper is based on nonlinear signaling rules and takes advantage of some useful properties of the Hopfield formulation. In particular, by considering two attractor states we will show that the network separates into two types of domains which do not interact with each other. Moreover, the Hopfield framework allows for a direct mapping of a gene expression pattern into an attractor state of the signaling dynamics, facilitating the integration of genomic data in the modeling.

The paper is structured as follows. In Section II we summarize the model and review some of its key properties. Section III describes general strategies aiming at selectively disrupting the signaling only in cells that are near a cancer attractor state. The strategies we have investigated use the concept of *bottlenecks*, which identify single nodes or strongly connected clusters of nodes that have a large impact on the signaling. In this section we also provide a theorem with bounds on the minimum number of nodes that guarantee controllability of a bottleneck consisting of a strongly connected component. This theorem is useful for practical applications since it helps to establish whether an exhaustive search for such minimal set of nodes is practical. In Section IV we apply the control strategies of Section III to lung and B cell cancers. We will use two different networks for this analysis. The first is an experimentally validated and non-specific network obtained from a kinase interactome and phospho-protein database [32] combined with a database of interactions between transcription factors and their target genes [33]. The second network is cell-specific and was obtained using network reconstruction algorithms and transcriptional and post-translational data from mature human B cells [34]. The algoritm reconstructed network is significantly more dense than the experimental one, and the same control strategies produce different results in the two cases. Conclusions are in Section V.

II. MODEL

We define the adjacency matrix of a network G as

$$A_{ij} = \begin{cases} 1 & \text{if } j \to i \\ 0 & \text{otherwise} \end{cases} , \tag{1}$$

where $j \to i$ denotes a directed edge from node j to node i. The set of nodes in the network G is indicated by $V(G)$ and the set of ordered pairs by $E(G) = \{(j,i) : j \to i\}$. In our analysis we assume that G is given. The spin of node i at time t is $\sigma_i(t) = \pm 1$, and indicates an expresssed $(+1)$ or not expressed (-1) gene. We encode an arbitrary attractor state $\vec{\xi} = (\xi_1, \xi_2, ..., \xi_N)$, with $\xi_i = \pm 1$ by defining the coupling matrix

$$J_{ij} = A_{ij}\xi_i\xi_j \ . \tag{2}$$

The total field at node i is then $h_i = h_i^{\text{ext}} + \sum_j J_{ij}\sigma_j$, where h_i^{ext} is the external field applied to node i, which will be discussed below. The discrete-time update scheme is defined as

$$\sigma_i(t + \Delta t) = \begin{cases} +1 & \text{with prob. } (1 + \exp[-h_i(t)/T])^{-1} \\ -1 & \text{with prob. } (1 + \exp[+h_i(t)/T])^{-1} \end{cases} \tag{3}$$

where $T \geq 0$ is an effective temperature. For the remainder of the paper, we consider the case of $T = 0$ so that $\sigma_i = \text{sign}(h_i)$, and the spin is chosen randomly from ± 1 if $h_i = 0$. For convenience, we take $t \in \mathbb{Z}$ and $\Delta t = 1$. Nodes can be updated synchronously, and synchronous updating can lead to limit cycles in our model [9]. Nodes can also be updated separately and in random order (anynchronous updating), which does not result in limit cycles. All results presented in this paper use the synchronous update scheme.

Some nodes may have no incoming connections. According to Eq. (3), these *source nodes* flip randomly between $+1$ and -1. The sources are thus fixed to their initial states so that $\sigma_q(t) = \sigma_q(0)$ for all $q \in Q$, where Q is the set of source nodes. The source nodes flip if directly targeted by an external field. Biologically, genes at the "top" of a network are assumed to be controlled by elements outside of the network.

In application, two attractors are needed. Define these states as $\vec{\xi}^n$ and $\vec{\xi}^c$, the *normal state* and *cancer state*, respectively. The magnetization along attractor state a is

$$m^a(t) = \frac{1}{N}\sum_{i=1}^{N} \sigma_i(t)\xi_i^a \ , \tag{4}$$

where N is the number of nodes in the network. Note that if $m^a(t) = \pm 1$, $\vec{\sigma}(t) = \pm\vec{\xi}^a$. We also define the steady state magnetization along state a as

$$m_\infty^a = \lim_{\tau\to\infty} \frac{1}{\tau}\sum_{t=1}^{\tau} m^a(t) \ . \tag{5}$$

There are two ways to model normal and cancer cells. One way is to simply define a different coupling matrix for each attractor state a,

$$J_{ij}^a = A_{ij}\xi_i^a\xi_j^a \ . \tag{6}$$

Alternatively, both attractor states can be encoded in the same coupling matrix,

$$J_{ij} = A_{ij}(\xi_i^n\xi_j^n + \xi_i^c\xi_j^c) \ . \tag{7}$$

Systems using Eqs. 6 and 7 will be referred to as the one attractor state ($p = 1$) and two attractor state ($p = 2$) systems, respectively. Eqs. 6 and 7 are particular cases of the general Hopfield form

$$J_{ij} = A_{ij}\sum_{k=1}^{p}\xi_i^k\xi_j^k \ , \tag{8}$$

where p is the number of attractor states, often taken to be large. An interesting property emerges when $p = 2$, however. Consider a simple network composed of two nodes, with only one edge $1 \to 2$ with attractor states $\vec{\xi}^n$ and $\vec{\xi}^c$, and $T = 0$. The only nonzero entry of the matrix J_{ij} is

$$J_{21} = \xi_2^n\xi_1^n + \xi_2^c\xi_1^c \ . \tag{9}$$

Note that if $\vec{\xi}^n = \pm\vec{\xi}^c$, $J_{21} = 2\xi_2^n\xi_1^n$. In either case, by Eq. (3) we have

$$\sigma_2(t+1) = \begin{cases} +\xi_2^n & \text{if } \sigma_1(t) = +\xi_1^n \\ -\xi_2^n & \text{if } \sigma_1(t) = -\xi_1^n \end{cases} \ , \tag{10}$$

that is, the spin of node 2 at a given time step will be driven to match the attractor state of node 1 at the previous time step. However, if $\xi_1^n = \pm\xi_1^c$ and $\xi_2^n = \mp\xi_2^c$, $J_{21} = 0$. This gives

$$\sigma_2(t) = \begin{cases} +1 & \text{with probability } 1/2 \\ -1 & \text{with probability } 1/2 \end{cases} \tag{11}$$

In this case, node 2 receives no input from node 1. Nodes 1 and 2 have become effectively disconnected.

This motivates new designations for node types. We define *similarity nodes* as nodes with $\xi_i^n = \xi_i^c$, and *differential nodes* as nodes with $\xi_i^n = -\xi_i^c$. We also define the set of similarity nodes $S = \{i : \xi_i^n = \xi_i^c\}$ and the set of differential nodes $D = \{i : \xi_i^n = -\xi_i^c\}$. Connections between two similarity nodes or two differential nodes remain in the network, whereas connections that link nodes of different types transmit no signals. The effective deletion of edges between nodes means that the original network fully separates into two subnetworks: one composed entirely of similarity nodes (the *similarity network*) and another composed entirely of differential nodes (the *differential network*), each of which can be composed of one or more separate weakly connected components (see Fig. 1). With this separation, new source nodes (*effective sources*) can be exposed in both the similarity and differential networks. For the remainder of this article, Q is the set of both source and effective source nodes in a given network.

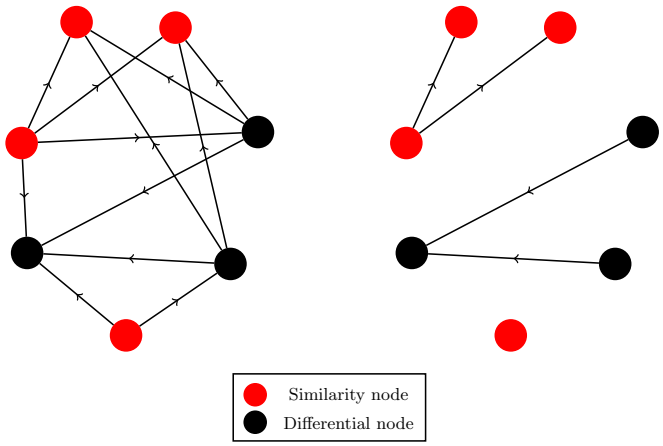

FIG. 1. For $p = 2$, every edge that connects a similarity node to a differential node or a differential node to a similarity node transmits no signal. Thus, the signaling in the right network shown above is identical to that of the left network. Because the goal is to leave normal cells unaltered while damaging cancer cells as much as possible, all similarity nodes can be safely ignored, and searches and simulations only need to be done on the differential subnetwork.

III. CONTROL STRATEGIES

The optimal choice of control strategy depends on the control goals, the network topology, the effective topology created by the attractor states, and the set of directly controllable nodes. The strategies presented below focus on selecting the best single nodes or small clusters of nodes to control, ranked by how much they individually change m_∞^a. In application, however, controlling many nodes is necessary to achieve a sufficiently changed m_∞^a. The effects of controlling a set of nodes can be more than the sum of the effects of controlling individual nodes, and predicting the truly optimal set of nodes to target is computationally difficult. Here, we discuss heuristic strategies for controlling large networks where the combinatorial approach is impractical.

For both $p = 1$ and $p = 2$, simulating a cancer cell means that $\vec{\sigma}(0) = +\vec{\xi}^c$, and likewise for normal cells. Although the normal and cancer states are mathematically interchangeable, biologically we seek to decrease m_∞^c as much as possible while leaving $m_\infty^n \approx +1$. By "network control" we thus mean driving the system away from its initial state of $\vec{\sigma}(0) = \vec{\xi}^c$ with $\vec{h}^{\text{ext}}$. Controlling individual nodes is achieved by applying an infinitely strong field to a set of targeted nodes T so that

$$h_\tau^{\text{ext}} = \begin{cases} \lim_{(u\to\infty)} -u\xi_\tau^c & \tau \in T \\ 0 & \text{else} \end{cases} \ . \tag{12}$$

This ensures that the drug field can always overcome the field from neighboring nodes.

In application, similarity nodes are never deliberately directly targeted, since changing their state would adversely affect both normal and cancer cells. Roughly

70% of the nodes in the networks surveyed are similarity nodes, so the search space is reduced. For $p = 2$, the effective edge deletion means that only the differential network in cancer cells needs to be simulated to determine the effectiveness of $\vec{h}^{\text{ext}}$. For $p = 1$, however, there may be some similarity nodes that receive signals from upstream differential nodes. In this case, the full effect of $\vec{h}^{\text{ext}}$ can be determined only by simulating all differential nodes as well as any similarity nodes downstream of differential nodes. All following discussion assumes that all nodes examined are differential, and therefore targetable, for both $p = 1$ and $p = 2$. The existence of similarity nodes for $p = 1$ only limits the set of targetable nodes.

A. Directed acyclic networks

Full control of a directed acyclic network is achieved by forcing $\sigma_q = -\xi_q^c$ for all $q \in Q$. This guarantees $m_\infty^c = -1$. Suppose that nodes $q \in Q$ in an acyclic network have always been fixed away from the cancer state, that is, $\sigma_q(t \to -\infty) = -\xi_q^c$. For any node i to have $\sigma_i(t) = \xi_i^n$, it is sufficient to have either $i \in Q$ or $\sigma_j(t-1) = \xi_j^n$ for all $j \to i$, $i \notin Q$. Because there are no cycles present, all upstream paths of sufficent length terminate at a source. Because the spin of all nodes $q \in Q$ point away from the cancer attractor state, all nodes downstream must also point away from the cancer attractor state. Thus, for acyclic networks, forcing $\sigma_q = -\xi_q^c$ guarantees $m_\infty^c = -1$. The complications that arise from cycles are discussed in the next subsection. However, controlling the original and effective sources may not be the most efficient way to push the system away from the cancer basin of attraction and, depending on the control limitations, it may not be possible. If minimizing the number of controllers is required, searching for the most important bottlenecks is a better strategy.

Consider a directed network G and an initially identical copy, $G' = G$. If removing node i (and all connections to and from i) from G' decreases the indegree of at least one node $j \in V(G')$, $j \neq i$, to less than half of its indegree in network G, $\{i\}$ is a *size 1 bottleneck*. The *bottleneck control set* of bottleneck $\{i\}$, $L(i)$, is defined algorithmically as follows: (1) Begin a set $L(i)$ with the current bottleneck i so that $L = \{i\}$; (2) Remove bottleneck $\{i\}$ from network G'; (3) Append $L(i)$ with all nodes j with current indegree that is less than half of that from the original network G; (4) Remove all nodes j from the network G'. If additional nodes in G' have their indegree reduced to below half of their indegree in G, go to step 3. Otherwise, stop. The *impact of the bottleneck* i, $I(i)$, is defined as

$$I(i) = |L(i)| \ , \tag{13}$$

where $|X|$ is the cardinality of the set X. The impact of a bottleneck is the minimum number of nodes that are guaranteed to switch away from the cancer state when the bottleneck is forced away from the cancer state.

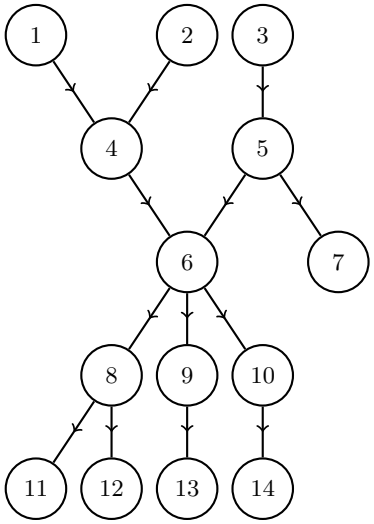

FIG. 2. An acyclic network. Controlling all three source nodes (nodes 1, 2 and 3) guarantees full control of the network, but are ineffective when targeted individually. The best single node to control in this network is node 6 because it directly controls all downstream nodes.

The impact is used to rank the size 1 bottlenecks by importance, with the most important as those with the largest impact. In application, when searching for nodes to control, any size 1 bottleneck $\{i\}$ that appears in the bottleneck control set of a different size 1 bottleneck $\{j\}$ can be ignored, since fixing j to the normal state fixes i to the normal state as well. Note that the definition given above in terms of G and G' avoids miscounting in the impact of a bottleneck. The network in Fig. 2, for example, has three sources (nodes 1, 2 and 3), but one important bottleneck (node 6). If full damage, i.e. $m_\infty^c = -1$, is required, then control of all source nodes is necessary. If minimizing the number of directly targeted nodes is important and $m_\infty^c > -1$ can be tolerated, then control of the bottleneck node 6 is a better choice.

B. Directed cycle-rich networks

Not all networks can be fully controlled at $T = 0$ by controlling the source nodes, however. If there is a cycle present, paths of infinite length exist and the final state of the system may depend on the initial state, causing parts of the network to be hysteretic. Controlling only sources in a general directed network thus does not guarantee $m_\infty^c = -1$ unless the system begins with $\sigma_i = -\xi_i^c$.

Define a *cycle cluster*, C, as a strongly connected subnetwork of a network G. The network in Fig. 3, for example, has one cycle cluster with nodes $V(C) = \{4, 5, 6, 7\}$. If the network begins with $\vec{\sigma}(0) = \vec{\xi}^c$, forcing both source nodes away from the cancer state does nothing to the

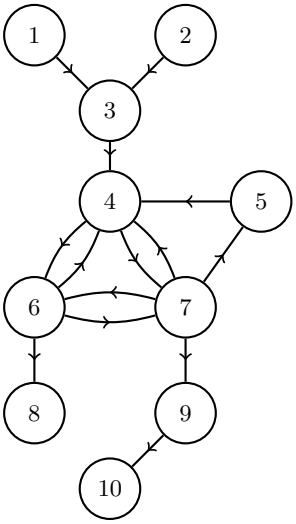

FIG. 3. A network with a cycle cluster composed of nodes 4, 5, 6 and 7. The high connectivity of node 4 prevents any changes made to the spin of nodes 1-3 from propagating downstream. The only way to indirectly control nodes 8-10 is to target nodes inside of the cycle cluster. Targeting node 4, 6 or 7 will cause the entire cycle cluster to flip away from its initial state, guaranteeing control of nodes 4-10 (see Fig. 4).

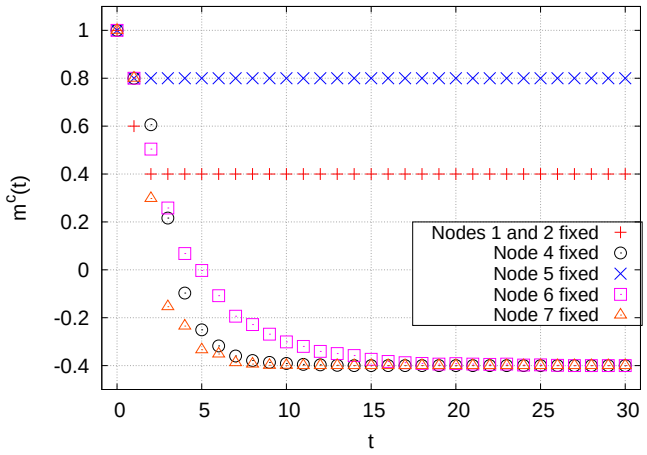

FIG. 4. Cancer magnetization from targeting various nodes in the network shown in Fig. 3, averaged over 10,000 runs. The averaging removes fluctuations due to the random flipping of nodes with $h_i = 0$. Targeting node 7 results in the quickest stabilization, but targeting any one of nodes 4, 6 or 7 results in the same final magnetization.

nodes downsteam of node 3 (see Fig. 4). This is because the indegree $\deg^-(4) = 4$, and a majority of the nodes connecting to node 4 are in the cancer attractor state. At $T = 0$, cycle clusters with high connectivity tend to block incoming signals from outside of the cluster, resulting in an insurmountable activation barrier.

The most effective single node to control in this network is any one of nodes 4, 6 or 7. Forcing any of these away from the cancer attractor state will eventually cause the entire cycle cluster to flip away from the cancer state, and all nodes downstream will flip as well, as shown in Fig. 4. The cycle cluster here acts as a sort of large, hysteretic bottleneck. We now generalize the concept of bottlenecks.

Define a *size k bottleneck* in a network G to be a cycle cluster B with $|V(B)| = k$ which, when removed from G, reduces the indegree of at least one node $j \in V(G)$, $j \notin V(B)$ to less than half of its original indegree. Other than now using the set of nodes $V(B)$ rather than a single node set, the above algorithm for finding the bottleneck control set remains unchanged. In Fig. 3, for instance, $V(B) = \{4, 5, 6, 7\}$, $k = 4$, $L(B) = \{4, 5, 6, 7, 8, 9, 10\}$, and $I(B) = 7$. With this more general definition, we note that controlling any size k bottleneck B guarantees control of all size 1 bottlenecks B' in the control set of B for all $k \geq 1$.

For any bottleneck B of size $k \geq 1$ in a network G, define the *set of critical nodes*, $Z(B, G)$, as the set of nodes $Z(B, G) \subseteq V(B)$ of minimum cardinality that, when controlled, guarantees full control of all nodes $i \in V(B)$. Also define the *critical number of nodes* as $n_{\text{crit}}(B, G) = |Z(B, G)|$. Thus, for the network in Fig. 4, $Z(B, G) = \{4\}$, $\{6\}$, or $\{7\}$, and $n_{\text{crit}}(B, G) = 1$.

In general, however, more than one node in a cycle cluster may need to be targeted to control the entire cycle cluster. Fig. 5 shows a cycle cluster (composed of nodes 2-10) that cannot be controlled by targeting any single node. The precise value of n_{crit} for a given cycle cluster C depends on its topology as well as the edges connecting nodes from outside of C to the nodes inside of C, and finding $Z(C, G)$ can be difficult. We present a theorem that puts bounds on n_{crit} to help determine whether a search for $Z(C, G)$ is practical.

Theorem: Suppose a network G contains a cycle cluster C. Define the *set of externally influenced nodes*

$$R(C, G) = \{i \in V(C) : j \in V(G \setminus C), (j, i) \in E(G)\} , \tag{14}$$

the *set of intruder connections*

$$W(C, G) = \{(j, i) \in E(G) : i \in V(C), j \in V(G \setminus C)\} , \tag{15}$$

and the *reduced set of critical nodes*

$$Z_{\text{red}}(C, G) = Z(C, G \setminus W) . \tag{16}$$

If $N = |V(C)|$ and

$$\mu \equiv \min_{i \in V(C)} \deg^-(i) , \tag{17}$$

where $\deg^-(i)$ is computed ignoring intruder connections, then

$$\left\lceil \frac{\mu}{2} \right\rceil \leq n_{\text{crit}}(C, G) \leq \zeta , \tag{18}$$

where

$$\zeta \equiv \min \left(\left\lceil \frac{N}{2} \right\rceil + |R(C, G) \setminus Z_{\text{red}}(C, G)|, N \right) . \tag{19}$$

Proof: First, prove the lower limit of Eq. (18). Let C be a cycle cluster in a network G with $R(C,G) = \{\varnothing\}$. (A cycle cluster in a network with $|R(C,G)| > 0$ will have the same or higher activation barrier for any node in the cluster than the same cycle cluster in a network with $R = \{\varnothing\}$. Since we are examining the lower limit of Eq. (18), we consider the case with the lowest activation barrier. Any externally influenced nodes cause $n_{\rm crit}$ to either increase or remain the same.) For any node i to be able to flip away from the cancer state (although not necessarily remain there), we must have that $h_i = -a\xi_i^c$ for $a \geq 0$, meaning that at least half of the nodes upstream of i must point away from the cancer state. The node i requiring the smallest number of upstream nodes to be in the normal state is the node that satisfies $\deg^-(i) = \mu$. Controlling less than $\mu/2$ nodes will leave all uncontrolled nodes with a field in the cancer direction, and no more flips will occur. Thus,

$$n_{\rm crit} \geq \left\lceil \frac{\mu}{2} \right\rceil \ . \tag{20}$$

For the upper limit of Eq. (18), consider a complete *clique* on N nodes, $C = K_N$ (that is, $A_{ij} = 1$ for all $i, j \in V(K_N)$, including self loops) in a network G. First, let there be no connections to any nodes in C from outside of C so that $R(C,G) = \{\varnothing\}$. For odd N, forcing $(N+1)/2$ nodes away from the cancer state will result in the field

$$\sum_j J_{ij}\sigma_j = \left(\frac{N-1}{2} - \frac{N+1}{2}\right)\xi_i^c = -\xi_i^c \tag{21}$$

for all nodes i. After one time step, all nodes will flip away from the cancer state. For even N, forcing $N/2$ nodes away from the cancer state will result in the field

$$\sum_j J_{ij}\sigma_j = \left(\frac{N}{2} - \frac{N}{2}\right)\xi_i^c = 0 \tag{22}$$

for all nodes i. At the next time step, the unfixed nodes will pick randomly between the normal and cancer state. If at least one of these nodes makes the transition away from the cancer state, the field at all other nodes will point away from the cancer direction. The system will then require one more time step to completely settle to $\sigma_i = \xi_i^c$. Thus, we have that for $C = K_N$ in a network G with $R(C,G) = \{\varnothing\}$,

$$n_{\rm crit}(K_N, G) = \left\lceil \frac{N}{2} \right\rceil \ . \tag{23}$$

K_N with $\sigma_i(0) = \xi_i^c$ gives the largest activation barrier for any cycle cluster on N nodes with $R(C,G) = \{\varnothing\}$ to switch away from the cancer attractor state. A general cycle cluster C with any topology on N nodes with $R(C,G) = \{\varnothing\}$ in a network G will have $\deg^-(i) \leq N$ for all nodes i, and so we have the upper bound

$$n_{\rm crit}(C,G) \leq \left\lceil \frac{N}{2} \right\rceil \ , \tag{24}$$

thus proving Eq. (18) for the special case of $R(C,G) = \{\varnothing\}$.

Now consider a cycle cluster C on N nodes in a network G with $|R(C,G)| \geq 0$. Suppose all nodes in $Z_{\rm red}(C,G)$ are fixed away from the cancer state. By Eq. (24), $|Z_{\rm red}(C,G)| \leq \lceil N/2 \rceil$. For any node $i \in (R(C,G) \cap Z_{\rm red}(C,G))$, $\sigma_i(t \to \infty) = -\xi_i^c$ is guaranteed because it has already been directly controlled. Any node $i \in (R(C,G) \setminus Z_{\rm red}(C,G))$ has some incoming connections from nodes $j \notin V(C)$, and these connections could increase the activation barrier enough such that fixing $Z_{\rm red}(C,G)$ is not enough to guarantee $\sigma_i(t \to \infty) = -\xi_i^c$. To ensure that any node $l \in V(C)$ points away from the cancer state, it is sufficient to fix all nodes $i \in (R(C,G) \setminus Z_{\rm red}(C,G))$ as well as $Z_{\rm red}(C,G)$ away from the cancer state. This increases $n_{\rm crit}$ by at most $|R(C,G) \setminus Z_{\rm red}(C,G)|$, leaving

$$n_{\rm crit}(C,G) \leq \left\lceil \frac{N}{2} \right\rceil + |R(C,G) \setminus Z_{\rm red}(C,G)| \ . \tag{25}$$

$n_{\rm crit}$ can never exceed N, however, because directly controlling every node results in controlling C. We can thus say that

$$n_{\rm crit}(C,G) \leq \min\left(\left\lceil \frac{N}{2} \right\rceil + |R(C,G) \setminus Z_{\rm red}(C,G)|, N\right) \ . \tag{26}$$

Finally, combining the upper limit in Eq. (26) with the lower limit from Eq. (20) gives Eq. (18). ■

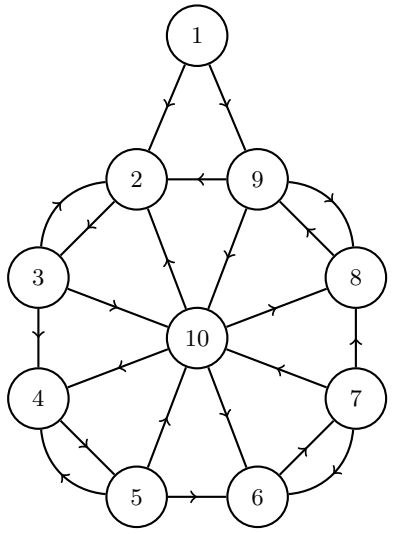

FIG. 5. A network with a cycle cluster C (composed of nodes 2-10) that cannot be controlled at $T = 0$ by controlling a single node. Here, $R(C,G) = \{2, 9\}$, $W(C,G) = \{(1,2), (1,9)\}$, $Z_{\rm red}(C,G) = \{9, 10\}$, $\mu = 1$ and $N = 9$, so $1 \leq n_{\rm crit} \leq 6$.

There can be more than one $Z_{\rm red}$ for a given cycle cluster. Note that the tightest constraints on $n_{\rm crit}$ in Eq. (18) come from using the $Z_{\rm red}$ with the largest overlap with R. If finding $Z_{\rm red}$ is too difficult, an overestimate

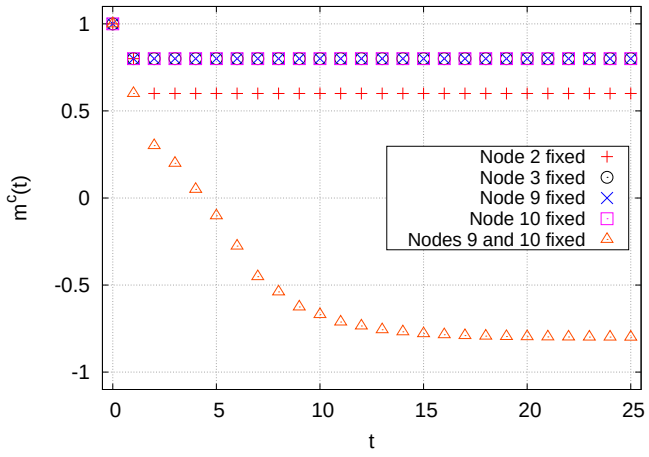

FIG. 6. Magnetization for network from Fig. 5, averaged over 10,000 runs. There is no single node to flip that will control the cycle cluster, but fixing nodes 9 and 10 results in full control of the cycle cluster, leaving only node 1 in the cancer state. This means $Z(C,G) = \{9, 10\}$ and $n_{\rm crit} = 2$.

for the upper limit of $n_{\rm crit}$ can be made by assuming that $R \cap Z_{\rm red} = \{\varnothing\}$ so that

$$\left\lceil \frac{\mu}{2} \right\rceil \leq n_{\rm crit}(C,G) \leq \min\left(\left\lceil \frac{N}{2} \right\rceil + |R(C,G)| \ , \ N \right) \ . \tag{27}$$

The cycle cluster in Fig. 5 has $N = 9$, $R = \{2, 9\}$, $\mu = 1$, and one of the reduced sets of critical nodes is $Z_{\rm red} = \{9, 10\}$, so $1 \leq n_{\rm crit} \leq 6$. It can be shown through an exhaustive search that for this network $n_{\rm crit} = 2$, and the set of critical nodes is $Z = \{9, 10\}$ (see Fig. 6). Here, $Z = Z_{\rm red}$, although this is not always the case. Because the cycle cluster has 9 nodes and $1 \leq n_{\rm crit} \leq 6$, at most $\sum_{n=1}^{6} \binom{9}{n} = 465$ simulations are needed to find at least one solution for $Z(C,G)$. However, the maximum number of simulations required to find $Z(C,G)$ increases exponentially and for larger networks the problem quickly becomes intractable.

One heuristic strategy for controlling cycle clusters is to look for size $k' < |V(C)|$ bottlenecks inside of C. Bottlenecks of size $k \gg 1$ and average indegree $\langle \deg^-(B) \rangle \ll k$ can contain high impact size k' bottlenecks, where $k' < k$. Size $k \geq 1$ bottlenecks need to be compared to find the best set of nodes to target to reduce m^c_∞. Simply comparing the impact is insufficent because a cycle cluster with a large impact could also have a large $n_{\rm crit}$, requiring much more effort than its impact merits. Define the *critical efficiency* of a bottleneck B as

$$e_{\rm crit}(B) = \frac{I(B)}{n_{\rm crit}(B,G)} \ . \tag{28}$$

If the critical efficiency of a cycle cluster is much smaller than the impacts of size 1 bottlenecks from outside of the cycle cluster, the the cycle cluster can be safely ignored.

For some cycle clusters, however, not all of the nodes need to be controlled in order for a large portion of the nodes in the cycle cluster's control set to flip. Define the *optimal efficiency* of a bottleneck B as

$$e_{\rm opt}(B) = \max_{n=1,2,\ldots} \left(\frac{I\left(\bigcup_{i=1}^{n} B_i\right)}{n} \right) \tag{29}$$

where $B_i \in V(B)$ are size 1 bottlenecks and $I(B_i) > I(B_{i+1})$ for all i. Note that for any size 1 bottleneck B, $e_{\rm opt}(B) = e_{\rm crit}(B) = I(B)$. This quantity thus allows bottlenecks with very different properties ($I(B)$, $n_{\rm crit}(B,G)$, or $|V(B)|$) to be ranked against each other.

All strategies presented above are designed to select the best individual or small group of nodes to target. Significant changes in the biological networks' magnetization require targeting many nodes, however. Brute force searches on the effect of larger combinations of nodes are typically impossible because the required number of simulations scales exponentially with the number of nodes. A crude Monte Carlo search is also numerically expensive, since it is difficult to sample an appreciable portion of the available space. Our alternative is to take advanatge of the bottleneck nodes that can be easily found, and rank all size $k \geq 1$ bottlenecks B_i in an ordered list U such that

$$U = (B_1, B_2, B_3, \ldots) \tag{30}$$

where

$$e_{\rm opt}(B_i) \geq e_{\rm opt}(B_{i+1}) \ , \ B_i \not\subset L(B_j) \tag{31}$$

for all $B_i, B_j \in U$ and fix the bottlenecks in the list in order. This is called the *efficiency-ranked* strategy. If all size $k > 1$ bottlenecks are ignored, it is called the *pure* efficiency-ranked strategy, and if size $k > 1$ bottlenecks are included it is called the *mixed* efficiency-ranked strategy.

An effective polynomial-time algorithm for finding the top z nodes to fix, which we call the *best+1* strategy (equivalent to a greedy algorithm), works as follows: (1) Begin with a seed set of nodes to fix, F; (2) Test the effect of fixing $F \cup i$ for all allowed nodes $i \notin F$; (3) $F \leftarrow F \cup i_{\rm best}$, where $i_{\rm best}$ is the best node from all i sampled; (4) If $|F| < z$, go to step (2). Otherwise, stop. The seed set of nodes could be the single highest impact size 1 bottleneck in the network, or it could be the best set of n nodes (where $n < z$) found from a brute force search.

IV. CANCER SIGNALING

In application to biological systems, we assume that the magnetization of cell type a is related to the *viability* of cell type a, that is, the fraction of cells of type a that survives a drug treatment. It is reasonable to assume

Properties	Lung	B cell
Nodes	9073	4364
Edges	45635	55144
Sources	129	8
Sinks	8443	1418
Av. outdegree	5.03	12.64
Max outdegree	240	2372
Max indegree	68	196
Self-loops	238	0
Undirected edges	350	23386
Diameter	11	11
Max cycle cluster	401	2886
Av. clustering coeff. [35]	0.0544	0.2315

TABLE I. General properties of the full networks. The network used for the analysis of lung cancer is a generic one obtained combining the data sets in Refs. [32] and [33]. The B cell network is a curated version of the B cell interactome obtained in Ref. [34] using a network reconstruction method and gene expression data from B cells.

	Lung		B					
Prop	I/A	I/H	N/D	N/F	N/L	M/D	M/F	M/L
N	1175	1320	885	724	1035	791	636	921
E	35821	33962	43471	40589	34791	38386	42030	35528

TABLE II. Nodes (N) and edges (E) of the differential networks obtained in the p=2 case after deleting similarity nodes for different pairs of normal and cancer attractor states. I = IMR-90 (normal), A = A549 (cancer), H = NCI-H358 (cancer), N = Naïve (normal), M = Memory (normal), D = DLBCL (cancer), F = Follicular lymphoma (cancer), L = EBV-immortalized lymphoblastoma (cancer).

that the viability of cell type a, $v^a(m^a_\infty)$, is a monotonically increasing function of m^a_∞. Because the exact relationship is not known, we analyze the effect of external perturbations in terms of the final magnetizations.

We need to use as few controllers as possible to sufficiently reduce m^c_∞ while leaving $m^n_\infty \approx +1$. In practical applications, however, one is limited in the set of druggable targets. All classes of drugs are constrained to act only on a specific set of biological components. For example, one class of drugs that is currently under intense research is protein kinase inhibitors [36]. In this case one has two constraints: the only nodes that can be targeted are those that correspond to kinases, and they can only be inhibited, i.e. turned off. We will use the example of kinase inhibitors to show how controllability is affected by such type of constraints. In the real systems studied, many differential nodes have only similarity nodes upstream and downstream of them, while the remaining differential nodes form one large cluster. This is not important for $p = 1$, but the effective edge deletion for $p = 2$ results in many *islets*, which are nodes i with $A_{ij} = A_{ji} = 0$ for all $i \neq j$ (self-loops allowed). Controlling islets requires targeting each islet individually. For $p = 2$, we concentrate on controlling only the largest weakly connected differential subnetwork. All final magnetizations are normalized by the total number of nodes in the full network, even if the simulations are only conducted on small portion of the network.

A. Lung cell network

The network used to simulate lung cells was built by combining the kinase interactome from PhosphoPOINT [32] with the transcription factor interactome from TRANSFAC [33]. Both of these are general networks that were constructed by compiling many observed pairwise interactions between components, meaning that if $j \to i$, at least one of the proteins encoded by gene j has been directly observed interacting with gene i in experiments. This bottom-up approach means that some edges may be missing, but those present are reliable. Because of this, the network is sparse ($\sim 0.057\%$ complete, see Table I), resulting in the formation of many islets for $p = 2$. Note also that this network presents a clear hierarchical structure, characteristic of biological networks [37, 38], with many "sink" nodes [39] that are targets of transcription factors and a relatively large cycle cluster originating from the kinase interactome. In our signaling model, the IMR-90 cell line was used for the normal attractor state, and the two cancer attractor states examined were from the A549 (adenocarcinoma) and NCI-H358 (bronchioalveolar carcinoma) cell lines. The resulting magnetization curves for A549 and NCI-H358 are very similar, so the following analysis addresses only A549. Table II lists the number of nodes and edges of the differential networks obtained using gene expression from these cell lines. The full network contains 9073 nodes, but only 1175 of them are differential nodes in the IMR-90/A549 model. In the unconstrained $p = 1$ case, all 1175 differential nodes are candidates for targeting. Exhaustively searching for the best pair of nodes to control requires investigating 689725 combinations simulated on the full network of 9073 nodes. However, 1094 of the 1175 nodes are sinks (i.e. nodes i with outdegree $\deg^+(i) = 0$, ignoring self loops) and therefore have $I(i) = e_{\rm opt}(i) = 1$, which can be safely ignored. The search space is thus reduced to 81 nodes, and finding even the best triplet of nodes exhaustively is possible. Including constraints, only 31 nodes are differential kinases with $\xi^c_i = +1$. This reduces the search space at the cost of increasing the minimum achievable m^c_∞.

There is one important cycle cluster in the full network, and it is composed of 401 nodes. This cycle cluster has an impact of 7948 for $p = 1$, giving a critical efficiency of at least ~ 19.8, and $1 \leq n_{\rm crit} \leq 401$ by Eq. 27. The optimal efficiency for this cycle cluster is $e_{\rm opt} = 29$, but this is achieved for fixing the first bottleneck in the cluster. Additionally, this node is the highest impact size 1 bot-

	Lung		B					
Properties	I/A	I/H	N/D	N/F	N/L	M/D	M/F	M/L
Nodes	506	667	684	511	841	621	457	742
Edges	846	1227	2855	1717	3962	2525	1501	3401
Sources and effective sources	30	34	12	11	9	9	9	12
Sinks and effective sinks	450	598	286	198	369	275	204	333
Av. outdegree	1.67	1.84	4.17	3.36	4.71	4.07	3.28	4.58
Max outdegree	52	51	155	143	336	138	132	292
Max indegree	8	10	40	29	49	35	27	44
Self-loops	27	31	0	0	0	0	0	0
Undirected edges	0	4	1238	738	1468	1000	596	1214
Diameter	9	9	12	15	12	13	14	12
Max cycle cluster size	6	3	351	280	397	305	199	337
Av. clustering coeff	0.0348	0.0421	0.1878	0.1973	0.2446	0.1751	0.1935	0.2389

TABLE III. Properties of the largest weakly connected differential subnetworks for all cell types. I = IMR-90 (normal), A = A549 (cancer), H = NCI-H358 (cancer), N = Naïve (normal), M = Memory (normal), D = DLBCL (cancer), F = Follicular lymphoma (cancer), L = EBV-immortalized lymphoblastoma (cancer).

tleneck in the full network, and so the mixed efficiency-ranked results are identical to the pure efficiency-ranked results for the unconstrained $p = 1$ lung network. The mixed efficiency-ranked strategy was thus ignored in this case.

Fig. 7 shows the results for the unconstrained $p = 1$ model of the IMR-90/A549 lung cell network [40]. The unconstrained $p = 1$ system has the largest search space, so the Monte Carlo strategy performs poorly. The best+1 strategy is the most effective strategy for controlling this network. The seed set of nodes used here was simply the size 1 bottleneck with the largest impact. Note that best+1 works better than effeciency-ranked. This is becasue best+1 includes the synergistic effects of fixing multiple nodes, while efficiency-ranked assumes that there is no overlap between the set of nodes downstream from multiple bottlenecks. Importantly, however, the efficiency-ranked method works nearly as well as best+1 and much better than Monte Carlo, both of which are more computationally expensive than the efficiency-ranked strategy.

Fig. 8 shows the results for the unconstrained $p = 2$ model of the IMR-90/A549 lung cell network. The search space for $p = 2$ is much smaller than that for $p = 1$. The largest weakly connected differential subnetwork contains only 506 nodes (see Table III) , and the remaining differential nodes are islets or are in subnetworks composed of two nodes and are therefore unnecessary to consider. Of these 506 nodes, 450 are sinks. If limiting the search to differential kinases with $\xi_i^c = +1$ and ignoring all sinks, $p = 2$ has 19 possible targets. There is only one cycle cluster in the largest differential subnetwork, containing 6 nodes. Like the $p = 1$ case, the optimal efficiency occurs when targeting the first node, which is the highest impact size 1 bottleneck. Because the mixed efficiency-ranked strategy gives the same results as the

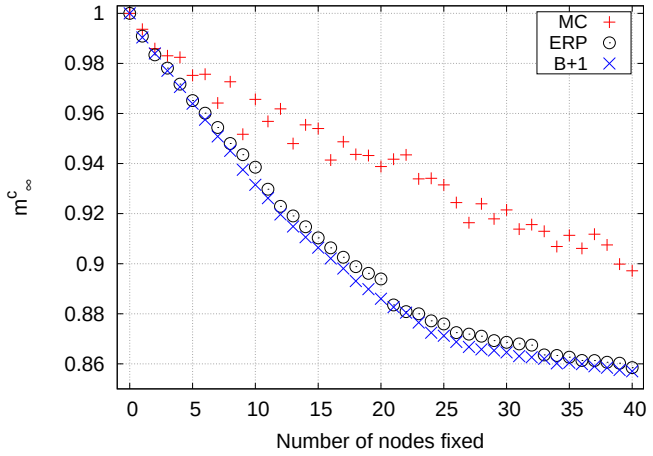

FIG. 7. Final cancer magnetizations for an unconstrained search on the lung cell network using $p = 1$. The efficiency-ranked strategy outperforms the relatively expensive Monte Carlo strategy. The best+1 strategy works best, although it requires the largest computational time. Note that the mixed efficiency-ranked curve is not shown because it is identical to the pure efficiency-ranked curve. Key: MC = Monte Carlo, B+1 = best+1, ERP = pure efficiency-ranked, ERM = mixed efficiency-ranked, EX = exhausive search.

pure efficiency-ranked strategy, only the pure strategy was examined. The Monte Carlo strategy fares better in the unconstrained $p = 2$ case because the search space is smaller. Additionally, the efficiency-ranked strategy does worse against the best+1 strategy for $p = 2$ than it did for $p = 1$. This is because the effective edge deletion decreases the average indegree of the network and makes nodes easier to control indirectly. When many upstream bottlenecks are controlled, some of the down-

stream bottlenecks in the efficiency-ranked list can be indirectly controlled. Thus, controlling these nodes directly results in no change in the magnetization. This gives the plateaus shown for fixing nodes 9-10 and 12-15, for example.

The only case in which an exhaustive search is possible is for $p = 2$ with constraints, which is shown in Fig. 9. Note that the polynomial-time best+1 strategy identifies the same set of nodes as the exponential-time exhaustive search. This is not surprising, however, since the constraints limit the available search space. This means that the Monte Carlo also does well. The efficiency-ranked method performs worst. The efficiency-ranked strategy is designed to be a heuristic strategy that scales gently, however, and is not expected to work well in such a small space when compared with more computationally expensive methods.

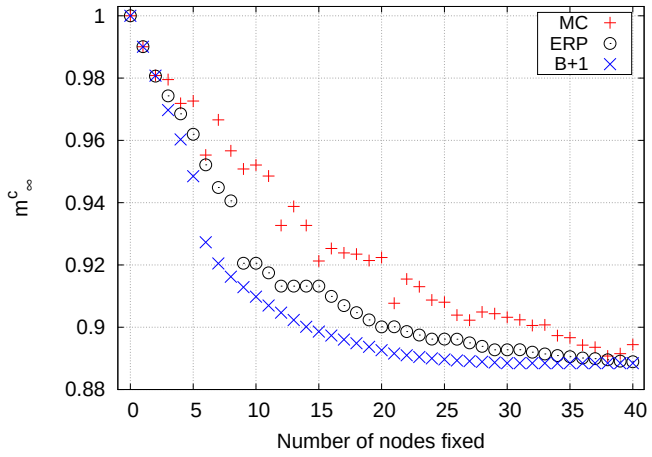

FIG. 8. Final cancer magnetizations for an unconstrained search on the lung cell network using $p = 2$. As in the $p = 1$ case, the efficiency-ranked strategy outperforms the expensive Monte Carlo search. The plateaus in the efficiency-ranked strategy when fixing 9-10, 12-15, 20-21, etc. nodes are a result of targeting bottlenecks that are already indirectly controlled.

B. B cell network

The B cell network was derived from the B cell interactome of Ref. [34]. The reconstruction method used in Ref. [34] removes edges from an initially complete network depending on pairwise gene expression correlation. Additionally, the original B cell network contains many protein-protein interactions (PPIs) as well as gene-gene interactions (GGIs). GGIs have definite directionality: a transcription factor encoded by one gene affects the expression level of its target gene(s). PPIs, however, do not have obvious directionality. We first filtered these PPIs by checking if the genes encoding these proteins interacted according to the PhosphoPOINT/TRANSFAC

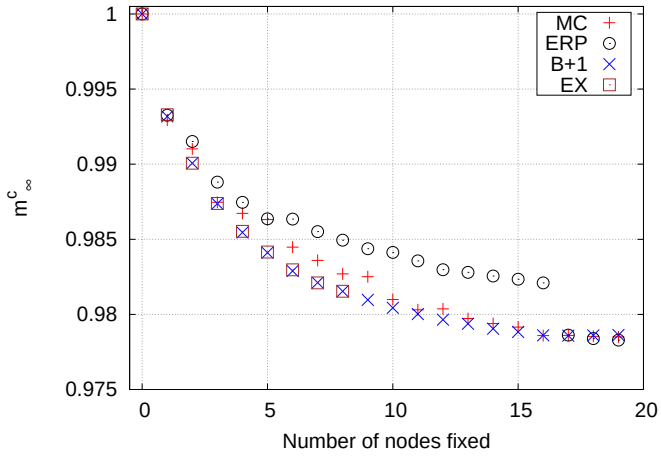

FIG. 9. Final cancer magnetizations for a constrained search on the lung cell network using $p = 2$. This is the only case in which a limited exhaustive search is possible. Interestingly, the exhaustive search locates the same nodes as the best+1 strategy for fixing up to eight nodes. The efficiency-ranked strategy performs poorly compared to the Monte Carlo strategy because the search space is small and a large portion of the available space is sampled by the Monte Carlo search.

network of the previous section, and if so, kept the edge as directed. If the remaining PPIs are ignored, the results for the B cell are similar to those of the lung cell network. We found more interesting results when keeping the remaining PPIs as undirected, as is discussed below.

Because of the network construction algorithm and the inclusion of many undirected edges, the B cell network is more dense (∼0.290% complete, see Table I) than the lung cell network. This higher density leads to many more cycles than the lung cell network, and many of these cycles overlap to form one very large cycle cluster containing ∼66% of nodes in the full network. We analyzed two types of normal B cells (naïve and memory) and three types of B cell cancers (diffuse large B-cell lymphoma (DLBCL), follicular lymphoma, and EBV-immortalized lymphoblastoma), giving six combinations in total. We present results for only the naïve/DLBCL combination below, but Tables II, III and IV list the properties of all normal/cancer combinations. The full B cell network is composed of 4364 nodes. For $p = 1$, there is one cycle cluster C composed of 2886 nodes. This cycle cluster has $1 \leq n_{\rm crit}(C) \leq 1460$, $I(C) = 4353$, and $3.0 \leq e_{\rm crit}(C) \leq 4353$. Finding $Z(C)$ was deemed too difficult.

Fig.10 shows the results for the unconstrained $p = 1$ case. Again, the pure efficiency-ranked strategy gave the same results as the mixed efficiency-ranked strategy, so only the pure strategy was analyzed. As shown in Fig. 10, the Monte Carlo strategy is out-performed by both the efficiency-ranked and best+1 strategies. The synergistic effects of fixing multiple bottlenecks slowly becomes

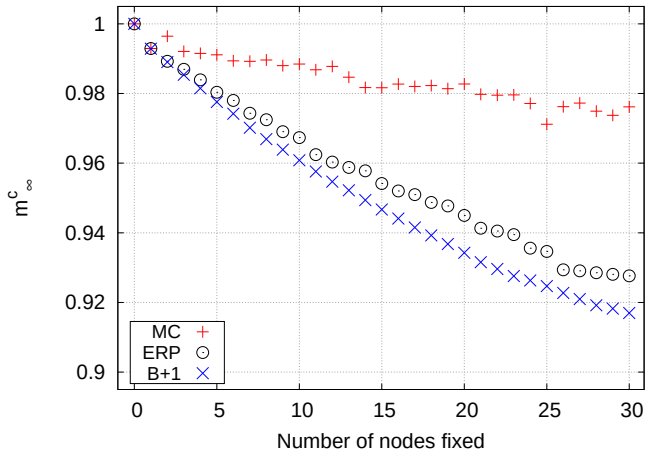

FIG. 10. Final cancer magnetizations for an unconstrained search on the B cell network using $p = 1$. The Monte Carlo strategy is ineffective for fixing any number of nodes. The efficiency-ranked and best+1 curves slowly separate because synergistic effects accumulate faster for best+1.

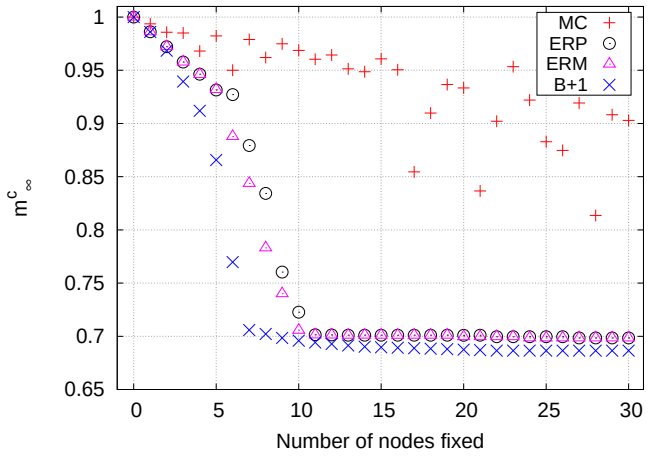

FIG. 11. Final cancer magnetizations for an unconstrained search on the B cell network using $p = 2$. The rather sudden drop in the magnetization between controlling 5 and 10 nodes in the efficiency-ranked strategies comes from flipping a significant portion of a cycle cluster. This is the only network examined in which the mixed efficiency-ranked strategy produces results different from the pure efficiency-ranked strategy.

apparent as the best+1 and efficiency-ranked curves separate.

Fig. 11 shows the results for the unconstrained $p = 2$ case. The largest weakly connected subnetwork contains one cycle cluster with 351 nodes, with $1 \leq n_{\rm crit} \leq 208$. Although finding a set of critical nodes is difficult, the optimal efficiency for this cycle cluster is 62.2 for fixing 10 bottlenecks in the cycle cluster. This makes targeting the cycle cluster worthwhile. The efficiency of this set of 10 nodes is larger than the efficiencies of the first 10 nodes from the pure efficiency-ranked strategy, so the m^c_∞ from the mixed strategy drops earlier than the pure strategy. Both strategies quickly identify a small set of nodes capable of controlling a significant portion of the differential network, however, and the same result is obtained for fixing more than 10 nodes. The best+1 strategy finds a smaller set of nodes that controls a similar fraction of the cycle cluster, and fixing more than 7 nodes results in only incremental decreases in m^c_∞. The Monte Carlo strategy performs poorly, never finding a set of nodes adequate to control a significant fraction of the nodes in the cycle cluster.

V. CONCLUSION

Signaling models for large and complex biological networks are becoming important tools for designing new therapeutic methods for complex diseases such as cancer. Even if our knowledge of biological networks is incomplete, fast progress is currently being made using reconstruction methods that use large amounts of publicly available omic data [12, 13]. The Hopfield model we use in our approach allows mapping of gene expression patters of normal and cancer cells into stored attractor states of the signaling dynamics in directed networks. The role of each node in disrupting the network signaling can therefore be explicitly analyzed to identify isolated genes or sets of strongly connected genes that are selective in their action. We have introduced the concept of *size k bottlnecks* to identify such genes. This concept led to the formulation of several heuristic strategies, such as the *efficiency-ranked* and *best+1* strategy to find nodes that reduce the overlap of the cell network with a cancer attractor. Using this approach, we have located small sets of nodes in lung and B cancer cells which, when forced away from their initial states with local magnetic fields (representing targeted drugs), disrupt the signaling of the cancer cells while leaving normal cells in their original state. For networks with few targetable nodes, exhaustive searches or Monte Carlo searches can locate effective sets of nodes. For larger networks, however, these strategies become too cumbersome and our heuristic strategies represent a feasible alternative. For tree-like networks, the pure efficiency-ranked strategy works well, whereas the mixed efficiency-ranked strategy could be a better choice for networks with high-impact cycle clusters.

Some of the genes identified in Table IV are consistent with current clinical and cancer biology knowledge. For instance, in the lung cancer list we found a well known tumor suppressor gene (TP53) [41] that is frequently mutated in many cancer types including lung cancer [42]. Mutations in PBX1 have recently been detected in non-small-cell lung cancer and this gene is now being consid-

	I/A				I/H			
	$p=1$		$p=2$		$p=1$		$p=2$	
	Gene	I	Gene	I	Gene	I	Gene	I
UNC	HNF1A	29	OR5I1	35	HNF1A	29	HMX1	41
	TMEM37	22	TMEM37	25	MAP3K3	18	PBX1	38
	OR5I1	20	HNF1A	23	TP53	18	MYB	25
	MAP3K14	19	POSTN	21	RUNX1	17	ITGB2	20
	MAP3K3	18	RORA	18	RORA	16	TNFRSF10A	18
CON	MAP3K14	19	SRC	15	TTN	16	BMPR1B	18
	SRC	14	BMPR1B	7	RIPK3	6	LCK	8

	N/D				N/F				N/L			
	$p=1$		$p=2$		$p=1$		$p=2$		$p=1$		$p=2$	
	Gene	I	Gene	I	Gene	I	Gene	I	Gene	I	Gene	I
UNC	BCL6	12	NFIC	22	BCL6	12	NCOA1	20	RBL2	11	RBL2	22
	MEF2A	5	TGIF1	19	MEF2A	5	NFATC3	15	FOXM1	8	ATF2	12
	NCOA1	5	BCL6	14	NCOA1	5	BCL6	11	ATF2	7	NFATC3	11
	TGIF1	4	FOXJ2	12	TGIF1	4	CEBPD	8	RXRA	5	RXRA	9
	NFATC3	4	NFATC3	12	NFATC3	4	RELA	8	NFATC3	4	PATZ1	8
CON	BUB1B	2	CSNK2A2	2	BUB1B	2	WEE1	2	BUB1B	2	PRKCD	2
	AAK1	1	AKT1	2	AAK1	1	CSNK2A2	2	AAK1	1	AURKB	2

	M/D				M/F				M/L			
	$p=1$		$p=2$		$p=1$		$p=2$		$p=1$		$p=2$	
	Gene	I	Gene	I	Gene	I	Gene	I	Gene	I	Gene	I
UNC	BCL6	12	FOXJ2	12	BCL6	12	NCOA1	18	RBL2	11	RBL2	16
	MEF2A	5	NFIC	12	MEF2A	5	BCL6	13	FOXM1	8	ATF2	10
	NCOA1	5	BCL6	11	NCOA1	5	E2F3	9	ATF2	7	ZNF91	8
	NFATC3	4	NCOA1	9	NFATC3	4	RUNX1	9	RXRA	5	STAT6	8
	SMAD4	4	MEF2A	8	RELA	4	TFE3	7	TGIF1	4	FOXM1	8
CON	AAK1	1	RIPK2	1	AAK1	1	ROCK2	2	AAK1	1	AURKB	2
	RIPK2	1	MAST2	1	RIPK2	1	RIPK2	1	SCYL3	1	RIPK2	1

TABLE IV. Best single genes and their impacts for the p=1 and p=2 models. The unconstrained (UNC) and constrained (CON) case are shown. The constrained case refer to target that are kinases and are expressed in the cancer case. I = IMR-90 (normal), A = A549 (cancer), H = NCI-H358 (cancer), N = Naïve (normal), M = Memory (normal), D = DLBCL (cancer), F = Follicular lymphoma (cancer), L = EBV-immortalized lymphoblastoma (cancer).

ered as a target for therapy and prognosis [43]. MAP3K3 and MAP3K14 are in the MAPK/ERK pathway which is a target of many novel therapeutic agents [44], and SRC is a well known oncogene and a candidate target in lung cancer [45]. BCL6 (B-cell lymphoma 6) is the most common oncogene in DLBCL, and it is known that its expression can predict prognosis and response to drug therapy [46]. BCL6 is also frequently mutated in follicular lymphoma [47]. Our analysis identified BCL6 as an important drug target for both DLBCL and follicular lymphoma using either naive or memory B-cells as a control for both $p = 1$ and $p = 2$. RBL2 disregulation has been recently associated with many types of lymphoma [48]. FOXM1 is a potential therapeutic target in mature B cell tumors [49] and ATF2 has been recently found to be highly disregulated in lymphoma [50]. Besides BCL6 discussed above, the N/D list for DLBC contains genes (MEF2A [51], NCOA1 [52], TGIF1 [53], NFATC3 [54]) that are all known to have a functional role in cancer, even if they have not been associated to the specific B-cell cancer types we have considered. Our predictions are for the immortalized cell lines we have selected, some of which are commonly used for in-vitro testing in many laboratories. RNAi and targeted drugs could then be used in these cell lines against the top scoring genes in Table IV to test the disruption of survival or proliferative capacity. If experimentally validated, our analysis based on attractor states and bottlenecks could be applied to

patient-derived cancer cells by integrating in the model patient gene expression data to identify patient-specific targets.

The above unconstrained searches assume that there exists some set of "miracle drugs" which can turn any gene "on" and "off" at will. This limitation can be patially taken into account by using constrained searches that limit the nodes that can be addressed. However, even the constrained search results are unrealistic, since most drugs directly target more than one gene. Inhibitors, for example, could target differential nodes with $\xi_i^c = -1$ and $\xi_i^n = +1$, which would damage only normal cells. Additionally, drugs would not be restricted to target only differential nodes, and certain combinations could be toxic to both normal and cancer cells. Few cancer treatments involve the use of a single drug, and the synergistic effects of combining multiple drugs adds yet another level of complication to finding an effective treatment [27]. On the other hand, the intrinsic nonlinearity of a cellular signaling network, with its inherent structure of attractor states, enhances controllability [31] so that a properly selected set of druggable targets might be sufficient for robust control.

ACKNOWLEDGMENTS

We would like to acknowledge support by the Congressionally Directed Medical Research Program (DOD) Lung Cancer Research (W81XWH-12-1-0233). GP and CP acknowledge support by Salgomed, Del Mar CA. We thank Andrew Hodges and Jacob Feala for help with biological datasets. Correspondence and requests for materials should be addressed to szedlak1@msu.edu.

[1] J.J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," Proc. Nat. Acad. Sci. USA **79**, 2554–2558 (1982).

[2] D.J. Amit, H. Gutfreund, and H. Sompolinsky, "Spin-glass models of neural networks," Phys. Rev. A **32**, 1007 (1985).

[3] B. Derrida, E. Gardner, and A. Zippelius, "An exactly solvable asymmetric neural network model," Europhys. Lett.) **4**, 167 (1987).

[4] A. H. Lang, H. Li, J. J. Collins, and P. Mehta, "Epigenetic landscapes explain partially reprogrammed cells and identify key reprogramming genes," ArXiv e-prints (2012), arXiv:1211.3133 [q-bio.MN].

[5] Ron C. Anafi and Jason H. T. Bates, "Balancing robustness against the dangers of multiple attractors in a hopfield-type model of biological attractors," PLoS ONE **5**, e14413 (2010).

[6] M. Aldana, S. Coppersmith, and L.P. Kadanoff, "Boolean dynamics with random couplings," in *Perspectives and Problems in Nonlinear Sciences* (Springer, 2003) pp. 23–89.

[7] S.A. Kauffman, "Metabolic stability and epigenesis in randomly constructed genetic nets," J. Theor. Biol. **22**, 437 – 467 (1969).

[8] S. Amari, H. Ando, T. Toyoizumi, and N. Masuda, "State concentration exponent as a measure of quickness in kauffman-type networks," Phys. Rev. E **87**, 022814 (2013).

[9] T. Rohlf and S. Bornholdt, "Self-organized criticality and adaptation in discrete dynamical networks," in *Adaptive Networks* (Springer, 2009) pp. 73–106.

[10] K.E. Kürten, "Correspondence between neural threshold networks and kauffman boolean cellular automata," J. Phys. A **21**, L615 (1988).

[11] K.E. Kürten, "Critical phenomena in model neural networks," Phys. Lett. A **129**, 157 – 160 (1988).

[12] R. De Smet and K. Marchal, "Advantages and limitations of current network inference methods," Nature Rev. Microbiol. **8**, 717–729 (2010).

[13] A.J. Hartemink, "Reverse engineering gene regulatory networks," Nature Biotechnol. **23**, 554–555 (2005).

[14] L. Bullinger, K. Döhner, E. Bair, S. Fröhling, R.F. Schlenk, R. Tibshirani, H. Döhner, and J.R. Pollack, "Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia," New Engl. J. Med. **350**, 1605–1616 (2004).

[15] K. Eppert, K. Takenaka, E.R. Lechman, L. Waldron, B. Nilsson, P. van Galen, K.H. Metzeler, A. Poeppl, V. Ling, J. Beyene, *et al.*, "Stem cell gene expression programs influence clinical outcome in human leukemia," Nature Med. **17**, 1086–1093 (2011).

[16] S. Huang, G. Eichler, Y. Bar-Yam, and D.E. Ingber, "Cell fates as high-dimensional attractor states of a complex gene regulatory network," Phys. Rev. Lett. **94**, 128701 (2005).

[17] H. Sui, I. Ernberg, and S. Kauffman, "Cancer attractors: A systems view of tumors from a gene network dynamics and developmental perspective," Sem. Cell Dev. Biol. **20**, 869 – 876 (2009).

[18] P. Ao, D. Galas, L. Hood, and X. Zhu, "Cancer as robust intrinsic state of endogenous molecular-cellular network shaped by evolution," Med. Hypotheses **70**, 678–684 (2008).

[19] J.D. Feala, J. Cortes, P.M. Duxbury, C. Piermarocchi, A.D. McCulloch, and G. Paternostro, "Systems approaches and algorithms for discovery of combinatorial therapies," Wiley Interdisciplinary Reviews: Systems Biology and Medicine **2**, 181–193 (2010).

[20] P. Creixell, E. M Schoof, J.T. Erler, and R. Linding, "Navigating cancer network attractors for tumor-specific therapy," Nature Biotechnol. **30**, 842–848 (2012).

[21] D. Calzolari, G. Paternostro, P.L. Harrington Jr., C. Piermarocchi, and P.M. Duxbury, "Selective control of the apoptosis signaling network in heterogeneous cell populations," PLoS ONE **2**, e547 (2007).

[22] V. Ágoston, P. Csermely, and S. Pongor, "Multiple weak hits confuse complex systems: a transcriptional regulatory network as an example," Phys. Rev. E **71**, 051909 (2005).

[23] Péter Csermely, Vilmos Ágoston, and Sandor Pongor, "The efficiency of multi-target drugs: the network ap-

proach might help drug design," Trends in Pharmacological Sciences **26**, 178–182 (2005).

[24] E.D. Sontag, *Mathematical control theory: deterministic finite dimensional systems*, Vol. 6 (Springer, 1998).

[25] T. Akutsu, M. Hayashida, W.K. Ching, and M.K. Ng, "Control of boolean networks: hardness results and algorithms for tree structured networks," J. Theor. Biol. **244**, 670–679 (2007).

[26] A. Choudhary, A. Datta, M.L. Bittner, and E.R. Dougherty, "Intervention in a family of boolean networks," Bioinformatics **22**, 226–232 (2006).

[27] J.D. Feala, J. Cortes, P.M. Duxbury, A.D. McCulloch, C. Piermarocchi, and G. Paternostro, "Statistical properties and robustness of biological controller-target networks," PLoS ONE **7**, e29374 (2012).

[28] N. Bhardwaj, M.B. Carson, A. Abyzov, K.-K. Yan, H. Lu, and M.B. Gerstein, "Analysis of combinatorial regulation: scaling of partnerships between regulators with the number of governed targets," PLoS Comp. Biol. **6**, e1000755 (2010).

[29] Yang-Yu Liu, Jean-Jacques Slotine, and Albert-László Barabási, "Controllability of complex networks," Nature **473**, 167–173 (2011).

[30] M. D. Plummer and L. Lovász, *Matching theory* (Elsevier, 1986).

[31] S.P. Cornelius, W.L. Kath, and A.E Motter, "Realistic control of network dynamics," Nature Commun. **4**, 1–9 (2013).

[32] C.-Y. Yang, C.-H. Chang, Y.-L. Yu, T.-C. E. Lin, S.-A. Lee, C.-C. Yen, J.-M. Yang, J.-M. Lai, Y.-R. Hong, T.-L. Tseng, K.-M. Chao, and C.-Y. F. Huang, "Phosphopoint: a comprehensive human kinase interactome and phospho-protein database," Bioinformatics **24**, i14–i20 (2008).

[33] V. Matys, E. Fricke, R. Geffers, E. Gössling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, *et al.*, "Transfac: transcriptional regulation, from patterns to profiles," Nucleic Acids Res. **31**, 374–378 (2003).

[34] C. Lefebvre, P. Rajbhandari, M.J. Alvarez, P. Bandaru, W.K. Lim, M. Sato, K. Wang, P. Sumazin, M. Kustagi, B.C Bisikirska, *et al.*, "A human b-cell interactome identifies myb and foxm1 as master regulators of proliferation in germinal centers," Mol. Syst. Biol. **6** (2010).

[35] G. Fagiolo, "Clustering in complex directed networks," Phys. Rev. E **76**, 026107 (2007).

[36] P. Cohen, "Protein kinases - the major drug targets of the twenty-first century?" Nature Rev. Drug Discov. **1**, 309–315 (2002).

[37] E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai, and A.-L. Barabási, "Hierarchical organization of modularity in metabolic networks," Science **297**, 1551–1555 (2002).

[38] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," Proc. Nat. Acad. Sci. USA **99**, 7821–7826 (2002).

[39] S.S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, "Network motifs in the transcriptional regulation network of escherichia coli," Nature Genet. **31**, 64–68 (2002).

[40] All simulations were performed using MATLAB on a desktop computer. Finding the efficiency-ranked curves took roughly one minute each, whereas the best+1 curves required between thirty minutes and several hours each. All points in the Monte Carlo curves report the smallest m^c_∞ computed from 200 trials for the given number of controllers. Each Monte Carlo curve required several hours. Running the simulations to make all curves in Sec. IV required approximately 12 hours in total.

[41] S.J. Baker, E.R. Fearon, J. M. Nigro, A.C. Preisinger, J.M. Jessup, D.H. Ledbetter, D.F. Barker, Y. Nakamura, R. White, B. Vogelstein, *et al.*, "Chromosome 17 deletions and p53 gene mutations in colorectal carcinomas," Science **244**, 217–221 (1989).

[42] T. Takahashi, M.M. Nau, I. Chiba, M.J. Birrer, R.K. Rosenberg, M. Vinocour, M. Levitt, H. Pass, A.F. Gazdar, and J.D. Minna, "p53: a frequent target for genetic abnormalities in lung cancer," Science **246**, 491–494 (1989).

[43] M.-L. Mo, Z. Chen, H.-M. Zhou, H. Li, T. Hirata, D.M. Jablons, and B. He, "Detection of e2a-pbx1 fusion transcripts in human non-small-cell lung cancer," J. Exp. Clin. Canc. Res. **32**, 29 (2013).

[44] C. Montagut and J. Settleman, "Targeting the raf–mek–erk pathway in cancer therapy," Canc. Lett. **283**, 125–134 (2009).

[45] S.I. Rothschild, O. Gautschi, E.B. Haura, and F.M. Johnson, "Src inhibitors in lung cancer: current status and future directions," Clin. Lung Canc. **11**, 238–242 (2010).

[46] C.P. Hans, D.D. Weisenburger, T.C. Greiner, R.D. Gascoyne, J. Delabie, G. Ott, H.K. Müller-Hermelink, E. Campo, R.M. Braziel, E. S. Jaffe, *et al.*, "Confirmation of the molecular classification of diffuse large b-cell lymphoma by immunohistochemistry using a tissue microarray," Blood **103**, 275–282 (2004); A. Rosenwald, G. Wright, W.C. Chan, J. M. Connors, E. Campo, R.I. Fisher, R.D. Gascoyne, H.K. Muller-Hermelink, E.B. Smeland, J. M. Giltnane, *et al.*, "The use of molecular profiling to predict survival after chemotherapy for diffuse large-b-cell lymphoma," New Engl. J. Med. **346**, 1937–1947 (2002); J.N. Winter, E.A. Weller, S.J. Horning, M. Krajewska, D. Variakojis, T.M. Habermann, R.I. Fisher, P.J. Kurtin, W.R. Macon, M. Chhanabhai, *et al.*, "Prognostic significance of bcl-6 protein expression in dlbcl treated with chop or r-chop: a prospective correlative study," Blood **107**, 4207–4213 (2006).

[47] A. Diaz-Alderete, A. Doval, F. Camacho, L. Verde, P. Sabin, R. Arranz-Saez, C. Bellas, C. Corbacho, J. Gil, M. Perez-Martin, *et al.*, "Frequency of bcl2 and bcl6 translocations in follicular lymphoma: relation with histological and clinical features," Leukemia Lymphoma **49**, 95–101 (2008); T. Akasaka, I.S. Lossos, and R. Levy, "Bcl6 gene translocation in follicular lymphoma: a harbinger of eventual transformation to diffuse aggressive lymphoma," Blood **102**, 1443–1448 (2003).

[48] L. Wang, S. Pal, and S. Sif, "Protein arginine methyltransferase 5 suppresses the transcription of the rb family of tumor suppressors in leukemia and lymphoma cells," Mol. Cell. Biol. **28**, 6262–6277 (2008); G. De Falco, E. Leucci, D. Lenze, P.P. Piccaluga, P.P. Claudio, A. Onnis, G. Cerino, J. Nyagol, W. Mwanda, C. Bellan, *et al.*, "Gene-expression analysis identifies novel rbl2/p130 target genes in endemic burkitt lymphoma cell lines and primary tumors," Blood **110**, 1301–1307 (2007); P.P. Piccaluga, G. De Falco, M. Kustagi, A. Gazzola, C. Agostinelli, C. Tripodo, E. Leucci, A. Onnis, A. Astolfi, M. R. Sapienza, *et al.*, "Gene expression analysis uncovers similarity and differences among burkitt lymphoma subtypes," *ibid.* **117**, 3596–3608 (2011).

[49] V.S. Tompkins, S.-S. Han, A. Olivier, S. Syrbu, T. Bair, A. Button, Laura Jacobus, Zebin Wang, Samuel Lifton, Pradip Raychaudhuri, *et al.*, "Identification of candidate b-lymphoma genes by cross-species gene expression profiling," PLoS ONE **8**, e76889 (2013).

[50] B.C. Valdez, A.R. Zander, G. Song, D. Murray, Y. Nieto, Y. Li, R.E. Champlin, and B.S. Andersson, "Synergistic cytotoxicity of gemcitabine, clofarabine and edelfosine in lymphoma cell lines," Blood Canc. J. **4**, e171 (2014); J. Walczynski, S. Lyons, N. Jones, and W. Breitwieser, "Sensitisation of c-myc-induced b-lymphoma cells to apoptosis by atf2," Oncogene (2013).

[51] X. Bai, L. Wu, T. Liang, Z. Liu, J. Li, D. Li, H. Xie, S. Yin, J. Yu, Q. Lin, *et al.*, "Overexpression of myocyte enhancer factor 2 and histone hyperacetylation in hepatocellular carcinoma," J. Canc. Res. Clinic. Oncol. **134**, 83–91 (2008).

[52] S. Fabris, L. Mosca, G. Cutrona, M. Lionetti, L. Agnelli, G. Ciceri, M. Barbieri, F. Maura, S. Matis, M. Colombo, *et al.*, "Chromosome 2p gain in monoclonal b-cell lymphocytosis and in early stage chronic lymphocytic leukemia," Am. J. Hemat. **88**, 24–31 (2013); Y. Zhang, C. Duan, C. Bian, Y. Xiong, and J. Zhang, "Steroid receptor coactivator-1: A versatile regulator and promising therapeutic target for breast cancer," J. Steroid Biochem. **138**, 17 (2013).

[53] R. Hamid and S.J. Brandt, "Transforming growth-interacting factor tgif regulates proliferation and differentiation of human myeloid leukemia cells," Mol. Oncol. **3**, 451–463 (2009); T.N. Libório, E. N. Ferreira, F. C. Aquino Xavier, D. M. Carraro, L. P. Kowalski, F. A. Soares, and F.D. Nunes, "Tgif1 splicing variant 8 is overexpressed in oral squamous cell carcinoma and is related to pathologic and clinical behavior," Oral Surg. Oral Med. **116**, 614–625 (2013); M.T. Bengoechea-Alonso and J. Ericsson, "Tumor suppressor fbxw7 regulates tgfβ signaling by targeting tgif1 for degradation," Oncogene **29**, 5322–5328 (2010).

[54] S. Z. Glud, A. B. Sörensen, M. Andrulis, B. Wang, E. Kondo, R. Jessen, L. Krenacs, E. Stelkovics, M. Wabl, E. Serfling, *et al.*, "A tumor-suppressor function for nfatc3 in t-cell lymphomagenesis by murine leukemia virus," Blood **106**, 3546–3552 (2005).